THE

ADJUSTMENT OF ERRORS

IN

PRACTICAL SCIENCE

BY

R. W. M. GIBBS

B.A. (Camb.), B.Sc. (Lond.)

AGRICULTURAL COLLEGE AND RESEARCH INSTI[TUTE]
No. 2248
11 OCT 1930
COIMBATORE

OXFORD UNIVERSITY PRESS
LONDON : HUMPHREY MILFORD
1929

OXFORD UNIVERSITY PRESS
AMEN HOUSE, E.C. 4
LONDON EDINBURGH GLASGOW
LEIPZIG NEW YORK TORONTO
MELBOURNE CAPETOWN BOMBAY
CALCUTTA MADRAS SHANGHAI
HUMPHREY MILFORD
PUBLISHER TO THE
UNIVERSITY

Printed in Great Britain

PREFACE

THIS book is the result of an attempt to simplify and condense into a readable form the gist of the Theory of Errors. It is primarily intended for the use of Research Students in Practical Science, but chapter V and parts of chapter IX should be useful also to candidates reading Mathematics for the Natural Science Preliminary Examination of Oxford and for similar examinations. At the same time there are hopes that the book may make some appeal to non-specialist readers who are interested in such subjects as Education and Sociology, in which the deductions are generally quite delusive unless the significance of the Mathematical data is comprehended. Such readers may, without serious loss, freely skip the whole of the purely technical Mathematics, which has been isolated in the Appendix.

The book has been largely inspired by Dr. G. M. B. Dobson, F.R.S., to whom my grateful acknowledgements are due. In particular I owe to him the *Percentage-error* treatment of the distribution of errors, as well as the broad outlook on the *Theory of Correlation*.

I must also acknowledge my indebtedness to Professor Georges Dreyer, F.R.S., and to Mr. R. L. Vollum, D.Phil., for much valuable information, and to the former for permission to use the results of his researches in Biology.

R. W. M. G.

September 1929.

CONTENTS

I. INTRODUCTION 7

II. PROBABILITY, CHANCE AND ODDS 12

III. THE NORMAL CURVE OF ERRORS 16

The Normal Equation, the Mean Error, the formulae for h, the Standard Deviation and the Points of Inflexion, the Probable Error, the Percentage-error curve and the 2-Y table, Odds against a particular error exceeding nP.

IV. AREA AND VOLUME DISTRIBUTION 26

The Equations, the Integral Equations, the formulae for h, Points of Inflexion and Maximum Points, Probable Errors.

V. THE LINE OF CLOSEST FIT, THE MOST PROBABLE POINT, ETC. 33

The Normal Equations, the Most Probable point of intersection, the straight line of closest fit, finding equations to curves of particular types from given sets of data. Exercises.

VI. THE CORRELATION COEFFICIENT 41

Perfect correlation, the Regression lines, the Regression Equations, Negative correlation, and Angles of Regression, $r = \sqrt{\tan \alpha \tan \beta}$, the Significance of r in terms of n, the Contingency table.

VII. PARTIAL CORRELATION, WEIGHTED OBSERVATIONS, THE OBSERVATION ERROR, AND QUALITY CORRELATION 56

Partial correlation, Weighted and grouped observations, the Observation Error, Quality correlation.

VIII. THE CORRELATION RATIOS 65

Failure of the Correlation coefficient, the Correlation Ratios.

IX. THE ERROR OF THE LAST FIGURE, THE ERROR OF THE INTERVAL, ERRORS OF FUNCTIONS . . 70

Errors of the sums of 2, 3, 4, and 5 quantities (due to the last figure or to the Interval), Relative Errors, Maximum

Errors, Maximum Errors of Functions, the Probable Error of the Sum or Difference of two quantities in terms of the separate Probable errors.

APPENDIX (Items given separately on pp. 79, 80) . . 81

LIST OF SYMBOLS 106

LIST OF FORMULAE 108

INDEX 111

All facts, formulae, and equations required for reference are numbered consecutively throughout the book on the right of the pages. References to the Appendix are indicated by Roman numerals.

INTRODUCTION

COIMBATORE

No measurement of any real thing can ever be correct, for the simple reason that no instrument is capable of infinitely small displacements and no human eye can detect infinitesimal separations. Errors are therefore inevitable.

The word *'error'* in its present sense is strictly technical and does not include *'mistakes'*. When an investigator writes 0·57 for 0·75 this is a mistake, with which no possible system could cope. Errors can be broadly divided into two classes: **Systematic** and **Accidental.**

Systematic errors are those which can be traced to some definite cause, such as a wrong setting of the reading pointer of an instrument, incorrect calibration, and so on. Under this heading are *personal* errors, which are entirely due to the mentality of the observers, some persistently tending to over-estimate and some to underestimate. When differences between readings are concerned, personal errors are self-rectifying. In other cases they can be estimated and allowed for.

Accidental errors are due to an indefinite number of unknown causes, over which the operator has no control.

Suppose that X is the true value of a quantity, that X' is the
observed value, and that $X' = X + x$. Then x is the **error** ; it 1
is the quantity that must be added algebraically to the *true* value to make the *observed* value.*

It is obvious that systematic errors can be subject to no general laws, and throughout this book when *'errors'* are referred to *'accidental errors'* are to be understood.

If a large number of measurements were made of the same quantity, and if the resulting errors were entirely accidental

* In Astronomy and in older books, the error is taken to be $X - X'$, but it is now more usual, and more in keeping with the popular idea, to use the definition given above.

and independent, what sort of distribution would one expect on *a priori* grounds? It might be reasonably supposed (i) that the mean value would be nearly equal to the true value, (ii) that small errors would be more numerous than large ones, and (iii) that the positive and negative errors would be grouped about the mean value symmetrically.

In order to test the validity of these assumptions for a special case, the following experiment, which the reader can easily perform by himself, was carried out. A thick black line was drawn on a sheet of paper, and the spaces to the right and left of it were marked off by lines parallel to the black line and half an inch apart. The paper was then placed on the floor and a pencil was dropped on to it, aim being taken every time for the central line. The pencil-marks after 300 shots were counted and tabulated according to the half-inch intervals into which they fell.

The number of shots falling in a particular interval is called the **frequency** of the shots in that interval. The results are 2
here tabulated :

Interval	$-3\frac{1}{2}''$ to $-4''$	$-3\frac{1}{2}''$ to $-3''$	$-3''$ to $-2\frac{1}{2}''$	$-2\frac{1}{2}''$ to $-2''$	$-2''$ to $-1\frac{1}{2}''$
Frequency	2	3	2	6	12
Interval	$-1\frac{1}{2}''$ to $-1''$	$-1''$ to $-\frac{1}{2}''$	$-\frac{1}{2}''$ to 0	0 to $\frac{1}{2}''$	$\frac{1}{2}''$ to $1''$
Frequency	23	47	52	45	46
Interval	$1''$ to $1\frac{1}{2}''$	$1\frac{1}{2}''$ to $2''$	$2''$ to $2\frac{1}{2}''$	$2\frac{1}{2}''$ to $3''$	$3''$ to $3\frac{1}{2}''$
Frequency	31	16	10	2	3

The corresponding error-frequency graph was constructed by taking the frequency of the shots in each interval as the height of the mid-ordinate of that interval. The graph is shown in fig. 1.

In the experiment just described, the position of 0 on the x-axis may be regarded as the true value of a quantity, and the

positions of the shots as measurements of it. It usually happens, however, that the true value is not known and all measurements have to be made from the mean value. If M is the mean value of a quantity, and M' the measured value, then if $M' = M + m$, m is the deviation. When the true value 3

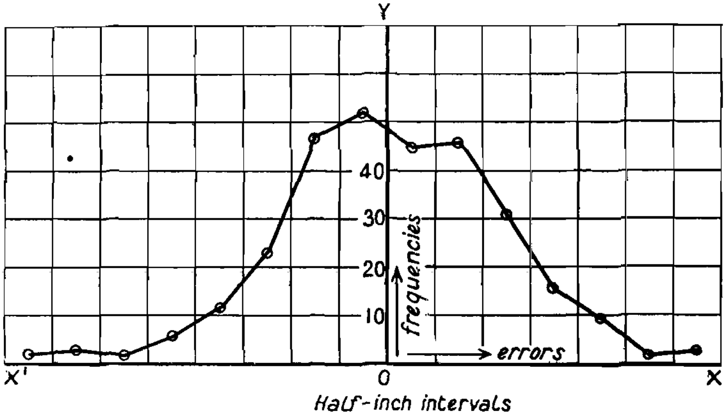


Fig. 1.

is not known, as is usually the case, the best assumption we can make is that M is the true value and that the deviations are true errors.

To illustrate this case a class of 42 boys were asked to measure the length (l) of a sheet of paper, in inches and decimals to the nearest hundredth of an inch.* The deviations from the Mean were taken as values of x, and the corresponding frequencies of occurrence as corresponding values of y'. The results were tabulated as follows:

l	8·60	8·59	8·58	8·57	8·56	8·55	8·54	8·53	8·52
x	0·037	0·027	0·017	0·007	−0·003	−0·013	−0·023	−0·033	−0·043
y'	2	3	4	9	10	8	4	1	1

Mean = 8·563 in.

* By the courtesy of Mr. J. A. Howells, Head Master of the South Oxford Boys' School.

The corresponding graph appears in fig. 2.

It will be noticed that the curve assumes a form similar to that of fig. 1.

The smallest quantity which an instrument can measure is
its **interval**. In fig. 1 the instrument was the set of parallel 4
lines and the interval was $\frac{1}{2}$ in.; in fig. 2 the instrument was
the ruler and the interval was $\frac{1}{100}$ in.

Figs. 1 and 2 differ in two particulars. In the former, the deviations were classified according to the intervals *in* which they fell. Moreover, to this degree of accuracy the deviations were true errors, because the position of the 'thick black line' was *definitely known*. In the latter, the deviations were assumed to fall exactly *on* the limits of the intervals, and the deviations were measured *from the Mean*.

Exercises.

1. Perform a similar experiment to that described on page 8.

2. Draw the error-frequency curve corresponding to the following set of estimates of a weight (w) in grammes; f is the frequency of each measurement. What is the Mean?

w	0·410	0·411	0·412	0·413	0·414	0·415	0·416	0·417	0·418
f	1	3	8	18	19	17	7	2	1

Ans. 0·4139.

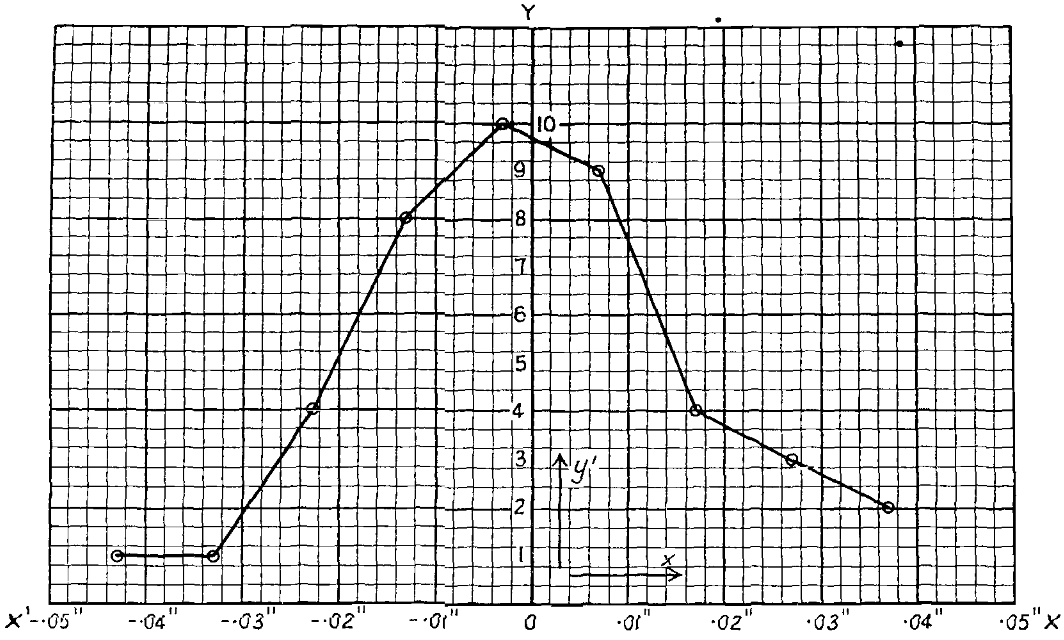


Fig. 2.

II

PROBABILITY, CHANCE AND ODDS

If an event can happen in a ways, and fail to happen in b ways, and if each of these ways of happening or not happening is equally likely, the **probability** or **chance** of its happening is $\frac{a}{a+b}$, and that of its *not* happening is $\frac{b}{a+b}$. In the case of a tossed penny, there are only 2 ways in which it can come down, 'heads' may happen in one way, and fail to happen in one way. Hence if we are considering the probability of heads turning up, $a = 1$, $b = 1$, and $\frac{a}{a+b} = \frac{1}{2}$, that is, the probability is $\frac{1}{2}$. **5**

In general, results based on the mathematical theory of probability are in agreement with experience. For example, if a penny is tossed up 100 times, it will invariably be found that heads will turn up *about* 50 times. If an event actually happens f out of n times, and if y is the probability of its happening, then if n is sufficiently large it may be assumed that $\mathrm{y} = \frac{\mathrm{f}}{\mathrm{n}}$, and that f (the frequency) $= \mathrm{yn}$ 6

Note that n in this formula corresponds to $(a+b)$ in the definition (5).

It is important to note, (i) that these results are only approximately true for any real case, and (ii) that they become more and more nearly true as n increases. In practice the assumption is always made that *frequencies are proportional to probabilities.*

Example 1. *What is the probability of picking a 'picture' card out of one suit?*

A 'picture' card can be drawn in 3 ways, and can fail to be drawn in 10 ways. Each of these ways is equally likely. Hence the probability of drawing a 'picture' card is $\frac{3}{13}$.

Example 2. *There are 2 white and 3 red balls in a bag. What is the chance of drawing a red ball? About how many times would one expect to draw a red ball if a ball were picked out 100 times, each time being replaced?*

The probability of drawing a red ball is $\frac{3}{5}$.

Since $f = yn$, where $y = \frac{3}{5}$ and $n = 100$.

$$\therefore \quad f = 100 \times \frac{3}{5} = 60.$$

The red ball would be drawn about 60 times.

Suppose that Jones tosses up a penny at the same time that Smith draws one card out of the 4 kings only. What is the chance that Jones throws a head while Smith draws the king of clubs?

Let us see what the possibilities are.

While Jones may throw

	a Head	or	*a Tail*
Smith may draw	Hearts	or	Hearts
	Clubs		Clubs
	Spades		Spades
	Diamonds		Diamonds

So there are 8 possibilities, and out of these, there is only **one** way in which Jones throws a tail and Smith draws the club. The probability of the double event is therefore $\frac{1}{8}$. In a very large number of cases, say 8,000 trials, *about* 1,000 cases would occur of tails and clubs. The larger the number of trials the more nearly would the double event come to $\frac{1}{8}$ of the total number.

Now the *probability* of Jones throwing a tail is $\frac{1}{2}$, and the *probability* of Smith drawing a club is 1 in 4 ($\frac{1}{4}$). So we see that the probability of the *double event* is the *product of the 2 separate probabilities.*

If the probability of one event happening is p_1 and that of another entirely independent event is p_2, then the probability of the two events happening simultaneously is $p_1 p_2$.

By extending the same reasoning it is not difficult to see that the probability (P) of n independent events happening

together, when the probabilities of the separate events are p_1, p_2, ... p_n is given by the formula:

$$\mathbf{P} = \mathbf{p}_1 \mathbf{p}_2 \mathbf{p}_3 \ldots \mathbf{p}_n . \qquad 7$$

(p_1 p_2 &c. are of course fractions.)

Example 3. *A and B both throw a die at the same time. What is the chance that they both throw a four?*

The chance of A throwing a four is $\frac{1}{6}$, and the chance of B throwing a four is $\frac{1}{6}$.

∴ The chance of both A and B throwing a four is $\frac{1}{36}$.

When the probability of an event is $\frac{a}{a+b}$, the odds in *favour* of it happening are $\frac{\mathbf{a}}{\mathbf{b}}$, and *against* it happening $\frac{\mathbf{b}}{\mathbf{a}}$. 8

If p is the probability of the event happening, and $d:1$ are the odds in favour of it happening, it is easy to show that $p = \frac{d}{d+1}$ and $d = \frac{p}{1-p}$. If $d':1$ are the odds *against* it happening, $p = \frac{1}{d'+1}$ and $d' = \frac{1-p}{p}$. 8

In example 3, the probability of the double event is $\frac{1}{36}$ and the *odds* against it happening are 35 to 1. If a man were to lay these odds he would lose 35 times and win once, in the long run, out of every 36 double throws. He would therefore neither lose nor gain.

Example 4. *If a penny is tossed up 4 times in succession, what are the odds against it turning up heads each time?*

Here each throw is entirely independent and it makes no difference that the throws are consecutive. The result is the same as if 4 different pennies were thrown up simultaneously.

The chance of one throw being a head is $\frac{1}{2}$.

∴ The chance of all 4 throws being heads is $(\frac{1}{2})^4 = \frac{1}{16}$.

The odds against are therefore 15 to 1.

Example 5. *The odds against a horse winning a race are 10 to 1. What is the chance that it will win?*

$$\text{Since } d' = \frac{10}{1}$$

$$\therefore \quad p = \frac{1}{d'+1} = \frac{1}{11}.$$

When an event is certain to happen, it is clear that the number of ways in which it can happen is equal to the number of ways in which it can or cannot happen, since $b = 0$ (5). In other words, **p** $= 1$ and *unity is the highest possible value of* **p**. 9

An interesting illustration of probability is the betting on a horse race. Since *one* horse is bound to win, the sum of the probabilities of the horses' winning ought to be 1. The following particulars of the betting on a race were picked at random out of a daily paper. There were 14 horses and the odds against were, 8 : 1, 8 : 1, 9 : 4, 100 : 14, 100 : 7, 100 : 6, and 8 horses at 33 : 1. The corresponding probabilities of winning were therefore estimated at $\frac{1}{9}$, $\frac{1}{9}$, $\frac{4}{13}$, $\frac{14}{114}$, $\frac{7}{107}$, $\frac{6}{106}$, and $\frac{1}{34}$. The sum of these probabilities is 1 : 1·010, which is surprisingly near to the theoretical value, unity.

The excess over unity is *against* the backer.

Exercises.

1. What is the probability that when 3 pennies are thrown up they will all come down 'tails'? What are the odds against this happening? *Ans.* $\frac{1}{8}$; 7 : 1.

2. If Jones is late at his office once a month, of 24 working days, and Smith is late once a week, of 6 working days, what are the odds against both being late the same morning? *Ans.* 143 : 1.

III

THE NORMAL CURVE OF ERRORS

THE experimental graphs of chapter I (figs. 1 and 2) would lead us to expect that the error-frequency curve in all cases would tend to assume the general shape foreshadowed by them. It may be stated at once that experience amply justifies the expectation.

Now this is an instance in which the pure mathematicians have been of the greatest assistance to scientists. The former by pure deductive reasoning have succeeded in finding *the equation to the curve.*

Let the errors arising from a large number (n) of observations be x_1, x_2, $x_3 \ldots x_n$. Suppose there is a common interval δx between each successive pair of errors and that a particular error x occurs y' times, so that y' is the frequency of x. If we assume that y' is a *continuous* function of x, we may regard y' as the frequency of errors lying between $x - \frac{1}{2}\delta x$ and $x + \frac{1}{2}\delta x$, or practically since δx is indefinitely small, between x and $x + \delta x$. The *probability* of a given error lying between x and $x + \delta x$ is $\frac{y'}{n}$ (Ref. 6.). If y be written for $\frac{y'}{n}$, it is shown in Appendix, II, that

$$y = \frac{h}{\sqrt{\pi}} e^{-h^2 x^2} dx. \quad \ldots\ldots\ldots\ldots$$

Before discussing this equation in detail attention is drawn to the following points.

In the practical graphs of chapter I, the heights of the ordinates represented frequencies instead of probabilities, but this is merely a question of scale (Ref. 6). It should be noted, however, that whereas the sum of the frequencies is equal to the number of measurements, the *sum of the probabilities is always equal to unity.* For since $y = \frac{f}{n}$ (Ref. 6),

$\Sigma(y) = \frac{\Sigma(f)}{n} = \frac{\text{Sum of frequencies}}{\text{Number of experiments}} = 1.$ The fact can also be looked at another way. The sum of the probabilities amounts to a certainty, and when the event is a certainty its probability is 1 (Ref. 9).

The **mean error** (M) of a set of measurements is the mean of all the errors *irrespective of sign*. **11**

The value of h for any set of experiments can be found in two ways, by the Standard method and by Peters's method. The two corresponding formulae are,

(i) $h = \sqrt{\frac{n}{2\Sigma(x^2)}}$. *Standard formula.** (V.) **12**

(ii) $h = \frac{1}{M\sqrt{\pi}}$, where M is the mean error. *Peters's formula.**

(VI.) **12a**

Taking the case of the 'school-boy' graph (fig. 2) we have the following figures: $n = 42$, $\Sigma(x^2) = 0{\cdot}0130$, and $M = 0{\cdot}014$. Using the standard formula, $h = 40{\cdot}2$, and from Peters's formula we have $h = 40{\cdot}3$. There is here an apparent discrepancy which needs a word of explanation.

Both of the formulae for h are based on equally sound theoretical grounds; there is no question of one being more reliable than the other. Both, however, are deduced from the assumption that the curve is *perfect*, that it is in fact the curve derived from the *equation*. From the nature of the case therefore the value of h can only be known approximately. In this case, the curve is far from perfect, and the errors themselves are only known to two-figure accuracy, on the assumption that $M = 8{\cdot}563$. The sum of the errors (irrespective of sign) and

* In these formulae the x's stand for *errors*, whereas in actual cases we are really dealing with *deviations*. Some writers attempt to allow for this unknown difference by replacing n by $(n-1)$, and n^2 by $n(n-1)$ in certain formulae. As, however, the formulae themselves are deduced from the assumption that the curves are perfect, which is never the case with experimental curves, the distinction serves no useful purpose. The alternative forms are, however, quoted in the Appendix.

the sum of the squares of the errors may be taken as expressed to three-figure accuracy. To this degree of accuracy the two values are in rough agreement. We shall take the value to be 40, for the sake of simplicity.

Since $h = 40$, and $\delta x = 0{\cdot}01$, we have $\frac{h}{\sqrt{\pi}}\,\delta x = \frac{0{\cdot}4}{\sqrt{\pi}} \triangleq 0{\cdot}23$.

Hence the equation to the theoretical curve is

$$y = 0{\cdot}23\,e^{-1600x^2}.$$

Here, y is the *probability* of an error. To compare the two curves it will be necessary to find the equation connecting the *frequencies* with the errors. This is easily done for $f = 42y$ (Ref. 6). Writing y' for f, we have

$$\therefore \quad y' = 9{\cdot}6\,e^{-1600x^2}.$$

The computation for the values of y' are here given.

x	x^2	$1600\,x^2$	$e^{1600\,x^2}$	$e^{-1600\,x^2}$	y'
0·00	0	0	1·00	1 00	9·6
±0·01	0·0001	0·16	1·17	0·85	8·2
±0·02	0·0004	0·64	1·90	0·53	5·1
±0·03	0·0009	1·44	4·22	0·24	2·3
±0·04	0·0016	2·56	12·9	0·08	0·8
±0·05	0·0025	4·00	54·6	0·02	0·2

In fig. 3 the practical graph and the theoretical one (broken line) are shown on the same frame of reference.*

The curve derived from the equation $y = \frac{h}{\sqrt{\pi}}\,e^{-h^2x^2}\,dx$ is the **Normal Curve of Errors.** It cannot be too clearly realized that the real and only value of the theoretical curve to the scientist is the fact that experiment does verify it, that it is, in fact, as its name suggests, the *Normal* curve. When a set of errors is *not* normally distributed, one is justified in assuming that there must be some special cause to account for the fact.

* If the more accurate value of h (40·25) is taken instead of 40, the resulting graph is almost indistinguishable from that of fig. 3.

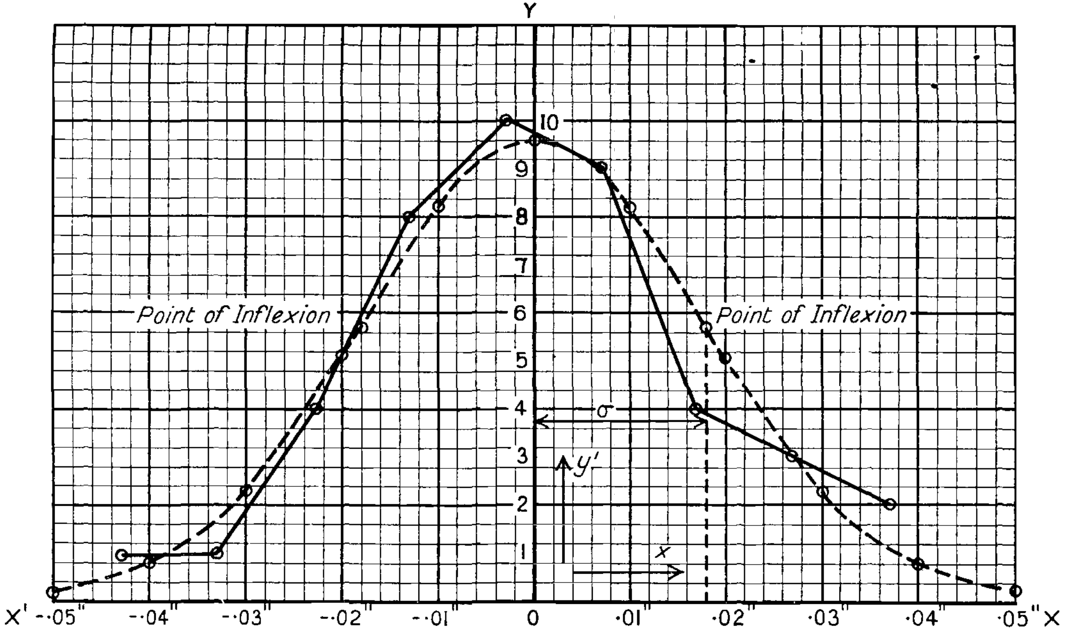

Y
10
9
8
7
6
5
4
3
2
1
Point of Inflexion
Point of Inflexion
σ
y'
x
X' -·05"
-·04"
-·03"
-·02"
-·01"
0
·01"
·02"
·03"
·04"
·05" X

FIG. 3.

There are two important and closely related functions in connexion with the normal curve which must be mentioned; the Standard Deviation, and the Probable Error.

The **standard deviation** of a set of errors is $\sqrt{\frac{\Sigma(x^2)}{n}}$, and is usually represented by the letter σ.

$$\sigma = \sqrt{\frac{\Sigma(x^2)}{n}} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad 1$$

The standard deviation is known also as the '*root-mean-square*' deviation, the '*virtual mean square*', and the '*radical-mean-square*'.

On the normal curve σ is the value of x, where the *point of inflexion* occurs (VII). In fig. 3, $\sigma = \sqrt{\frac{\Sigma(x^2)}{n}} = \sqrt{\frac{0{\cdot}0130}{42}} \triangleq 0{\cdot}018$, and the points of inflexion occur where $x = \pm 0{\cdot}018$. The standard deviation has also another significance unconnected with the curve; it may be regarded as a *measure of dispersion.* Scattered results give a high standard deviation, and results closely clustered about the mean value have a low one.

The **Probable Error** *is that error, on either the positive side or the negative one, which has as many errors numerically below as above it.* Because the normal curve is symmetrical the two values, which are only considered numerically, are assumed to be equal. 1

Suppose an error-frequency graph to be built up by taking the height of each mid-ordinate equal to the number of errors in the interval. Then since the intervals are all the same the *area* of each one must be proportional to the number of errors in it. In fig. 4, for example, the area enclosed by any two ordinates, the x-axis, and the broken-line top is proportional

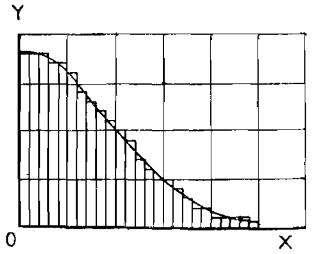


FIG. 4.

to the number of errors between those ordinates. But when the intervals are small the broken-line top merges into the smooth curve, and it may be said of the theoretical curve, where the intervals are infinitely small, that the area between any two ordinates is proportional to the number of errors falling between them. It follows that the ordinate through the *probable error bisects the area of one-half of the normal curve.*

It can be shown that if P is the probable error, then

$$P = \frac{0.4769}{h} \quad . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; 15$$

Taking the value of h for the 'school-boy' curve as 40, we see that the probable error is 0·012. If the areas on the positive side of the curve (fig. 3) below and above this error be added up in hundredths of a square inch, they will be found to be practically equal.

It must be pointed out that a value of P calculated from errors which are *not* normally distributed is meaningless. The value of P depends upon that of h, which is derived from the *normal equation.* Yet one frequently finds, even in standard books, probable errors worked out from five or six isolated experiments.

If h in formula 15 is replaced by its equivalents given in formulae 12 and 12a we have

$$P = 0.6745\sqrt{\frac{\Sigma(x^2)}{n}} = 0.6745\,\sigma \quad . \; . \; . \; . \; . \; 16$$

$$\text{or} \quad 0.8453\,\frac{\Sigma(x)}{n} = 0.8453\,M \quad . \; . \; . \; . \; . \; . \; 17$$

In words, these results are: '*The Probable Error is about* $\frac{2}{3}$ *of the Standard deviation or 85 per cent. of the Mean error*'. . . 18

The Probable Error (P_m) of the Mean of n observations is given by the formula,

$$P_m = \frac{P}{\sqrt{n}} \quad . \; . \; . \; . \; . \; . \; . \; . \; . \; . \; 19$$

The Percentage-Error Curve.

It is often more informative to know what percentage the sum of the errors not exceeding (numerically) a particular error

is of the whole number of errors, than it is to know what are the frequencies or probabilities of the errors themselves. To display this at a glance, an error percentage-error graph, or more shortly a 'percentage-error' graph, is constructed.

For an example, we will take the 'pencil-shot' curve of fig. 1. Taking the errors (x) irrespective of sign, the number of errors not exceeding $\frac{1}{2}$ in. was 97, or $32\frac{1}{3}$ per cent. of the whole number of errors (300), those not exceeding 1 in. were 190 or $63\frac{1}{3}$ per cent. of the total number, and so on. The graph is given in fig. 5, X being the number of errors not exceeding x.

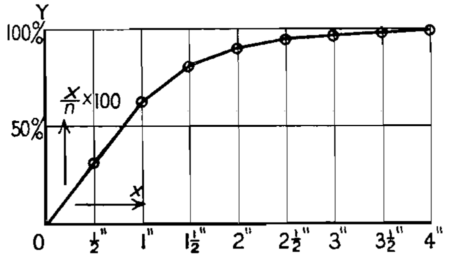


Fig. 5.

The probable error can be read off at once from the percentage-error curve, for it is only necessary to see where the 50-per-cent. line cuts it. In this case, the probable error is about $\frac{3}{4}$ in.

It was shown, in the paragraph leading up to formula 15, that the number of errors between any two ordinates on the error-frequency or the error-probability graph is proportional to the area between them. It follows that if the total area of the frequency graph is A, and the area up to a certain error x' is a, then the ordinate on the percentage-error curve for the error x' is $\frac{a}{A} \times 100$. The two curves are related as shown in the diagrams of fig. 6.

The percentage-error graph can be obtained in an exactly similar way from the error-probability graph.

Now the area of the error-probability curve from 0 to x' is $\frac{h}{\sqrt{\pi}}\int_0^{x'} e^{-h^2x^2}dx$, and the value of this integral from $-\infty$ to $+\infty$ is 1 (Appendix, I, or Ref. 9). So that $\frac{a}{A}\times 100$ is simply $100a$. It is usual to write Y instead of a, and to take the equation

$$\mathrm{Y} = \frac{\mathrm{h}}{\sqrt{\pi}}\int_0^{x} \mathrm{e}^{-\mathrm{h}^2\mathrm{x}^2}\,\mathrm{dx}$$

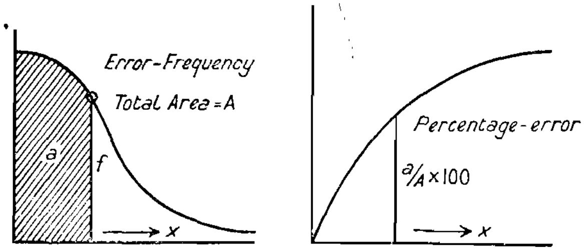


FIG. 6.

as the equation of the percentage-error curve, for whether we take Y or $100Y$ is merely a matter of adjusting a decimal point. The equation is also known as the *Integral Equation* for linear distribution.

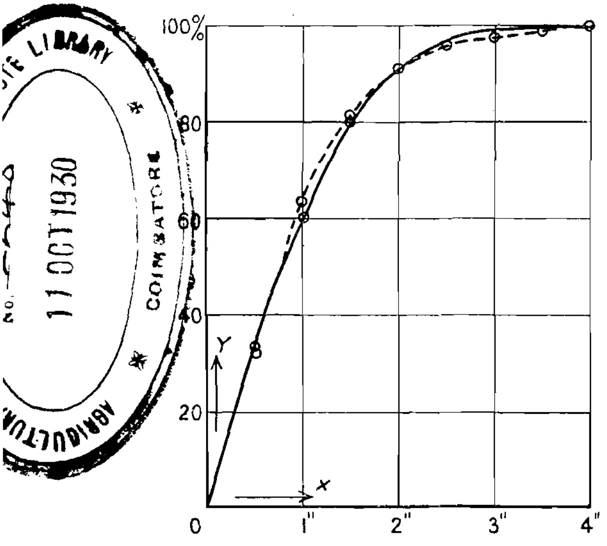
The integral $\int_0^{x} e^{-h^2x^2}dx$ cannot be found directly. It is handled practically as follows. By substituting z (say) for hx, we have

$$\frac{h}{\sqrt{\pi}}\int_0^{x'} e^{-h^2x^2}dx = \frac{1}{\sqrt{\pi}}\int_0^{hx'} e^{-z^2}dz = \frac{1}{\sqrt{\pi}}\int_0^{hx'} e^{-x^2}dx.$$

The latter function, which is equivalent to Y, can be evaluated numerically (Appendix, III). The table in Appendix, IV, gives the values of $2Y$ between the required limits. By taking $2Y$ instead of Y the limits are extended so that instead of being from 0 to hx' they are from $-hx'$ to $+hx'$. This enables one to deal with the positive and negative errors together. Suppose

for example we want Y for values of x not exceeding 0·01, on the 'school-boy' graph (taking the theoretical curve). Since $h = 40$, $hx = 0{\cdot}4$. From the Y-table we find corresponding to $hx = 0{\cdot}4$, $2\,Y = 0{\cdot}428$. That is, the errors lying between $-0{\cdot}01$ and $+0{\cdot}01$, or the errors not exceeding 0·01 numerically form 42·8 per cent of the whole. In this way we can easily build up the theoretical Y-curve.

The theoretical Y-curve for the 'pencil-shot' experiment is represented by the unbroken line in fig. 7. The dotted line shows the parts of the practically found curve where the two are not coincident.



LIBRARY
COIMBATORE
AGRICULTU
11 OCT 1930

FIG. 7.

The computation is as follows. First find the Mean error (M). To do this, consider all the shots in a compartment to lie on the mid-ordinate. The 97 shots between $-\frac{1}{2}''$ and $+\frac{1}{2}''$, for example, were taken as 97 shots at a distance of $\frac{1}{4}$ in. from the central line. M was found to be 0·95 in. approximately; from which $h \triangleq 0{\cdot}60$. Corresponding to $x = \frac{1}{2}''$, $1''$, $1\frac{1}{2}''$, &c., we have

$hx = 0{\cdot}3$, $0{\cdot}6$, $0{\cdot}9$, &c. The corresponding values of $2\,Y$ were then picked out from the $2{\cdot}Y$ table.

It is usual to think of the magnitude of errors in terms of the Probable Error or Standard Deviation. The Probable Error (corresponding to 50 per cent.) in fig. 7 is 0·8 in. The odds against a particular error exceeding 3 times the Probable Error are the odds against it exceeding 2·4. Now the percentage of errors exceeding 2·4 is 4, so that the odds against it are 96 : 4 or 24 : 1. The odds against a particular error exceeding nP or $n\sigma$, in both cases numerically, and with normal distribution, are given in the following tables:

nP	d'
P	1 : 1
$1\frac{1}{2}P$	2·2 : 1
$2P$	4·6 : 1
$3P$	22·3 : 1
$4P$	142 : 1
$5P$	1342 : 1
$6P$	19300 : 1 *

$n\sigma$	d'
σ	2·2 : 1
$1\frac{1}{2}\sigma$	6·5 : 1
2σ	21 : 1
3σ	369 : 1
4σ	15800 : 1 *

d'	nP	$n\sigma$
1 : 1	P	$0{\cdot}67\,\sigma$
5 : 1	$2{\cdot}05\,P$	$1{\cdot}38\,\sigma$
10 : 1	$2{\cdot}51\,P$	$1{\cdot}69\,\sigma$
10^2 : 1	$3{\cdot}82\,P$	$2{\cdot}58\,\sigma$
10^3 : 1	$4{\cdot}88\,P$	$3{\cdot}29\,\sigma$
10^4 : 1	$5{\cdot}77\,P$	$3{\cdot}89\,\sigma$
10^5 : 1	$6{\cdot}55\,P$	$4{\cdot}42\,\sigma$
10^6 : 1	$7{\cdot}25^*\,P$	$4{\cdot}89^*\,\sigma$

In the case of the 'school-boy' curve, taking the Probable error as 0·012 in., and the number of observations 42, the Probable Error of the Mean was $\frac{0{\cdot}012}{\sqrt{42}}$ in. (Ref. 19), which is approximately 0·00185 in. The odds against the error of the mean exceeding $5{\cdot}77\,P_m$ or 0·0107 in. numerically, is 10000 : 1 (roughly).†

* To 3-figure significance approximately.

† The Errors of the Means of sets of 42 measurements each, would be distributed normally like any other errors.

IV

AREA AND VOLUME DISTRIBUTION

[*It is not necessary to read this chapter to understand the rest of the book.*]

So far we have only considered errors as positive and negative lengths, or as distances to the right and left of a line. In dealing with the case of shots at a target, for example, the distribution of the errors considered as distances *from the centre* is quite different.

Let us take an actual case. A target was made by a series of concentric circles one inch apart, the bull's-eye being a small black spot at the centre. Eighty-one shots were made at the bull's-eye with the following results:

Distance from the centre	0 to 1″	1″ to 2″	2″ to 3″	3″ to 4″	4″ to 5″
Number of shots	22	38	14	6	1

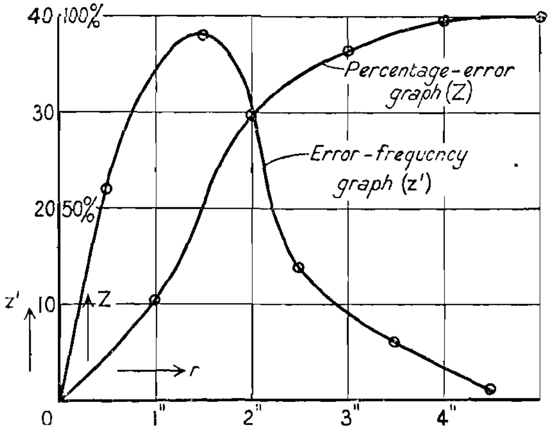


FIG. 8.
Graph from experiment.

The corresponding error-frequency (x, z') and percentage-error (x, Z) graphs are shown in fig. 8. Z is the percentage of

errors not exceeding r. For example, the errors not exceeding 2 in. were 60, and $\frac{60}{81} = 74$ per cent., so the ordinate corresponding to an error of 2 in. is 74 (see graph). The values of Z corresponding to 1″, 2″, 3″, 4″, and 5″, are 27 per cent., 74 per cent., 91 per cent., 99 per cent., and 100 per cent.

If z is the probability of the occurrence of an error (r), then

$$z = 2h^2 r e^{-h^2 r^2} dr \quad \text{(IX)} \quad . \quad . \quad . \quad . \quad . \quad . \quad \mathbf{21}$$

This is the equation of area distribution.

If Z is the area of the curve up to an error r, then

$$Z = 2h^2 \int_0^r r e^{-h^2 r^2} dr = 1 - e^{-h^2 r^2} \quad \text{(X)} \quad . \quad . \quad . \quad . \quad . \quad \mathbf{22}$$

This is the equation of the percentage-error graph (Ref. 20). It is also known as the *Integral Equation* of area distribution.

In area-distribution h can be expressed in terms of mean error (M):

$$h = \frac{\sqrt{\pi}}{2M} \quad \text{(XIII)} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad \mathbf{23}$$

This is the more convenient formula, but h can also be expressed in terms of the Standard Deviation (σ):

$$h = \frac{1}{\sigma} \quad \text{(XII)} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad \mathbf{24}$$

Both these formulae being deduced from the *ideal* curve, we cannot expect them to bring identical results from any *real* set of data. In this case the first formula yields $h = 0{\cdot}559$ and the second $0{\cdot}546$. For the purpose of drawing the graph, we shall take the value as $0{\cdot}55$.

In Appendix, XIX, it is shown that a maximum occurs where $r = \frac{1}{h\sqrt{2}}$, and a point of inflexion where $r = \sqrt{\frac{3}{2}} \cdot \frac{1}{h}$. **25**

Taking our value of $0{\cdot}55$ for h, it can easily be calculated that there should be a Maximum where $r = 1{\cdot}29$ and a point of inflexion where $r = 2{\cdot}23$. In the computation we will include the value $1{\cdot}3$ for r. This will verify the Maximum. The point of inflexion can be judged from the graph (fig. 9).

r	$2h^2r$	h^2r^2	$e^{h^2r^2}$	$e^{-h^2r^2}$	$\delta r = 1)$ $2h^2re^{-h^2r^2}\delta r$	$81z \equiv z'$
$\frac{1}{2}$	0·3025	0·0756	1·078	0·9276	0·2806	22·7
1·3	0·7865	0·5112	1·667	0·5999	0·4719	38·2
$1\frac{1}{2}$	0·9075	0·6806	1·975	0·5063	0·4595	37·2
$2\frac{1}{2}$	1·513	1·8906	6·623	0·1510	0·2285	18·5
$3\frac{1}{2}$	2·118	3·7056	40·67	0·0246	0·0521	4·2
$4\frac{1}{2}$	2·723	6·1256	457·4	0·0022	0·0060	0·5

The values of Z can be written down by inspection of the values of $e^{-h^2r^2}$ in the above table, the values corresponding to $r = \frac{1}{2}$, 1·3, &c., being, 7·2 per cent., 40·0 per cent., 49·4 per cent., 84·9 per cent., 97·5 per cent., and 99·8 per cent.

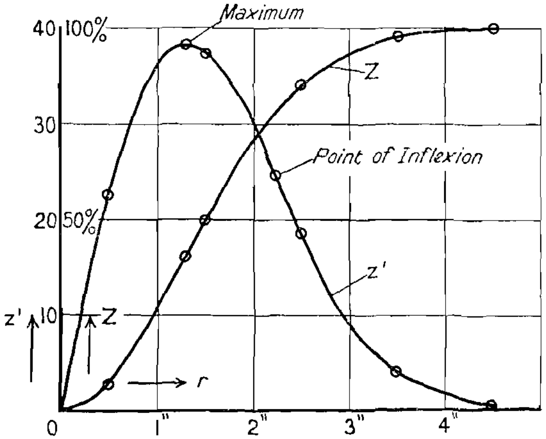


FIG. 9.

There is obviously a maximum value of z, where $r \doteq 1·3$, and a point of inflexion somewhere in the neighbourhood of $r = 2·2$.

The likeness between this curve and that of fig. 8 is really remarkable considering that the former was obtained from 81 random shots at a target.

The Probable Error (Q) in area distribution is given by the formula,

$$Q = \frac{0·8326}{h} \quad \text{(XI)} \; . \; . \; . \; . \; . \; . \; .$$

Expressing h by the values given in equations 24 and 23, we have respectively: $\mathbf{Q} = 0{\cdot}8326\sigma$ (XII) 27

or $\mathbf{Q} = 0{\cdot}9394\,\mathbf{M}$ (XIII) 28

In this case the first formula gives $Q = 1{\cdot}53$ ($\sigma = 1{\cdot}833$), and the second one, $Q = 1{\cdot}49$ ($M = 1{\cdot}586$). From the graph of Z, Q can be read off as about 1·51.

These connexions may be conveniently expressed in words. *The probable error is approximately 94 per cent. of the mean error or 83 per cent. of the standard deviation.* 29

The distribution of errors which are distances from a point in *space* is, as one would expect, different from linear and area distribution. This kind of distribution has to be considered in connexion with the molecular theory of gases. If v is the probability of the occurrence of an error s, then:

$$v = \frac{4h^3}{\sqrt{\pi}} s^2 e^{-h^2 s^2} ds \quad \text{(XIV)} \quad . \ . \ . \ . \ 30$$

The value of h in this case can be expressed in terms of the mean error (M) or the standard deviation (σ).

$$h = \frac{2}{M\sqrt{\pi}} \quad \text{(XVIII)} \quad . \ . \ . \ . \ . \ . \ . \ 31$$

$$\text{and} \quad h = \sqrt{\frac{3}{2}} \cdot \frac{1}{\sigma} \quad \text{(XVII)} \quad . \ . \ . \ . \ . \ . \ . \ 32$$

The Probable Error (R) in space distribution is given by the formula,

$$R = \frac{1{\cdot}0876}{h} \quad \text{(XVI)} \quad . \ . \ . \ . \ . \ . \ . \ 33$$

Replacing h. by the values given in 32 and 31, we have, respectively:

$$\mathbf{R} = 0{\cdot}8880\sigma$$

$$\text{or} \quad 0{\cdot}9639\mathbf{M}. \qquad 34$$

The probable error in space distribution is approximately 96 per cent. of the mean error or 89 per cent. of the standard deviation. 35

If V is the area of the space curve up to an error s, then

$$V = 2Y - \frac{2hs}{\sqrt{\pi}} e^{-h^2 s^2} \quad \text{(XV)} \quad . \ . \ . \ . \ . \ 36$$

where

$$Y = \frac{1}{\sqrt{\pi}} \int_0^{hx} e^{-x^2} dx \quad \text{(Ref. 20).}$$

Equation 36 is the equation of the percentage-error graph, also known as the *Integral Equation* of space distribution. (See p. 23.)

It is difficult to make experiments from which these curves can be constructed, but their forms can be studied from theoretically calculated values of v and V. For this purpose 10 and 0·05 have been taken as arbitrary values of h and δs respectively.

The computation is as follows:

s	s^2	$h^2 s^2$	$\left(\frac{4 h^3}{\sqrt{\pi}} \cdot s^2\right)$ $\frac{1000}{\sqrt{\pi}} s^2$	$\left(e^{h^2 s^2}\right)$ $e^{100 s^2}$	$e^{-100 s^2}$	$\delta s = \frac{1}{20}$ v	$200\ v \equiv v'$
0.05	0.0025	0.25	5.64	1.284	0.7788	0.2196	43.9
0.10	0.0100	1.00	22.56	2.718	0.3679	0.4150	83.0
0.15	0.0225	2.25	50.78	9.488	0.1054	0.2675	53.5
0.20	0.0400	4.00	90.24	54.60	0.0183	0.0826	16.5
0.25	0.0625	6.25	141.00	518.0	0.001931	0.0136	2.7
0.30	0.0900	9.00	203.04	8102	0.000123	0.0012	0.2

s	hs	$\left(\frac{2 hs}{\sqrt{\pi}}\right)$ $1.128\, s$	$\left(e^{-h^2 s^2}\right)$ $e^{-100 s^2}$	$\left(\frac{2 hs}{\sqrt{\pi}} e^{-h^2 s^2}\right)$ $1.128\, s\, e^{-100 s^2}$	$2V$	V	$100\ V$
0.05	0.5	0.5641	0.7788	0.4393	0.5205	0.0812	8.1
0.10	1.0	1.128	0.3679	0.4150	0.8427	0.4277	42.8
0.15	1.5	1.692	0.1054	0.1783	0.9661	0.7878	78.8
0.20	2.0	2.256	0.0183	0.0413	0.9953	0.9540	95.4
0.25	2.5	2.821	0.001931	0.005417	0.9996	0.9942	99.4
0.30	3.0	3.384	0.000123	0.000416	1.000	0.9996	100.0

The v is multiplied by 200 so as to make the figures more readily comparable with V. The fact of multiplying by 200 converts the error-probability curve into an error-frequency curve, when the total number of errors is 200. The two curves are shown in fig. 10.

The Mean of the errors comes to *0·1128*, from which the probable error can be calculated by formula 31. Its value

is 0·109, and this agrees very closely with the value as read off from the percentage-error graph.

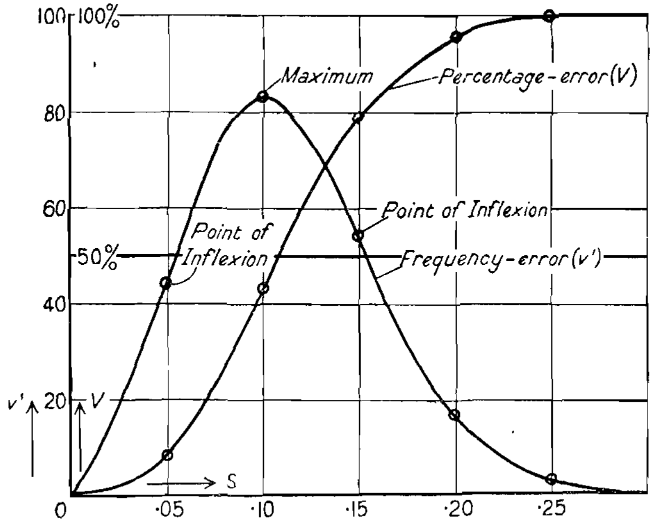


FIG. 10.

It is shown in Appendix. XX, that on the space-distribution curve there is a **Maximum** where $s = \frac{1}{h}$, and two points of **inflexion at** $s = \frac{\sqrt{5 \pm \sqrt{17}}}{2} \cdot \frac{1}{h}$, that is at $\frac{1{\cdot}5102}{h}$ and $\frac{0{\cdot}4682}{h}$.

On the particular curve we are dealing with, since $h = 10$, $\frac{1{\cdot}5102}{h}$ and $\frac{0{\cdot}4682}{h}$ come to 0·15 and 0·05 respectively, to 2-figure accuracy. Also when $s = \frac{1}{h} = 0{\cdot}1$ there is obviously a Maximum. So far as one can tell by eye, there are points of inflexion in the neighbourhood of 0·05 and 0·15. If it were desirable to obtain further arithmetical confirmation as to the point of inflexion at 0·05, for example, all that would be necessary would be to find v when $s = 0{\cdot}04$, 0·045, 0·055, and 0·060.

Exercise.

1. In a target there are 6 rings, counting the bull's-eye ring as one, and they are at equal distances apart. The following is the actual score made by a marksman in 208 shots:

Bull's-eye	1–2	2–3	3–4	4–5	5–6
41	115	39	12	1	0

Draw the error-frequency curve. Calculate the value of M, h, and Q. Check the value of Q by counting squares.

Ans. $M = 1{\cdot}62$, $h = 0{\cdot}547$, $Q = 1{\cdot}52$. The latter appears a little too low, but this is because on the theoretical curve the Max. occurs a little to the left of 1·5.

V

THE LINE OF CLOSEST FIT, THE MOST PROBABLE POINT, ETC.

SUPPOSE x and y are two quantities which *approximately* satisfy the three equations,

$$a_1x+b_1y+c_1=0$$
$$a_2x+b_2y+c_2=0$$
$$a_3x+b_3y+c_3=0.$$

In general, for any particular values of x and y, we shall have,

$$a_1x+b_1y+c_1=E_1$$
$$a_2x+b_2y+c_2=E_2$$
$$a_3x+b_3y+c_3=E_3$$

where E_1, E_2, and E_3 are the errors.

It has been deduced mathematically and verified by experience that the most probable values of x and y are those which produce values of E_1, E_2, and E_3, so that

$(E_1{}^2+E_2{}^2+E_3{}^2)$ is at a minimum.

Now $E_1{}^2=a_1{}^2x^2+b_1{}^2y^2+c_1{}^2+2a_1b_1xy+2a_1c_1x+2b_1c_1y$

$E_2{}^2=a_2{}^2x^2+b_2{}^2y^2+c_2{}^2+2a_2b_2xy+2a_2c_2x+2b_2c_2y$

$E_3{}^2=a_3{}^2x^2+b_3{}^2y^2+c_3{}^2+2a_3b_3xy+2a_3c_3x+2b_3c_3y.$

$\therefore\ E_1{}^2+E_2{}^2+E_3{}^2$

$$\equiv x^2\Sigma a^2+y^2\Sigma b^2+2xy\Sigma ab+2x\Sigma ac+2y\Sigma bc+\Sigma c^2.$$

Suppose we call this function u,

Then we know that:

$$\frac{\partial u}{\partial x}\,.\,dx+\frac{\partial u}{\partial y}\,.\,dy=du.$$

And if u is at a minimum, $du=0$.

This will be the case when $\frac{\partial u}{\partial x}=0$ and $\frac{\partial u}{\partial y}=0$.

Now $\frac{\partial u}{\partial x}=2x\Sigma a^2+2y\Sigma ab+2\Sigma ac.$

And $\frac{\partial u}{\partial y}=2y\Sigma b^2+2x\Sigma ab+2\Sigma bc.$

$\therefore$ The *best* or *most probable* values of x and y are obtained from the two equations,

$$x\Sigma a^2 + y\Sigma ab + \Sigma ac = 0$$

and $$x\Sigma ab + y\Sigma b^2 + \Sigma bc = 0. \qquad 38$$

These are the *Normal equations* for x and y respectively. A moment's consideration will be sufficient to show that the above reasoning would apply equally well to any number of equations.

To arrange the work for arithmetic computation, make three tables of a, b, c; a^2, b^2, ab, ac and bc as follows:

I.

a_1	b_1	c_1	σ_1
a_2	b_2	c_2	σ_2
a_3	b_3	c_3	σ_3
. . .	. . .	. . .	. . .
a_n	b_n	c_n	σ_n
Σa	Σb	Σc	$\Sigma \sigma$

where $\sigma = a + b + c$.

II.

a_1^2	$a_1 b_1$	$a_1 c_1$	$a_1 \sigma_1$
a_2^2	$a_2 b_2$	$a_2 c_2$	$a_2 \sigma_2$
a_3^2	$a_3 b_3$	$a_3 c_3$	$a_3 \sigma_3$
. . .	. . .	. . .	. . .
a_n^2	$a_n b_n$	$a_n c_n$	$a_n \sigma_n$
Σa^2	Σab	Σac	$\Sigma a\sigma$

III.

$a_1 b_1$	b_1^2	$b_1 c_1$	$b_1 \sigma_1$
$a_2 b_2$	b_2^2	$b_2 c_2$	$b_2 \sigma_2$
$a_3 b_3$	b_3^2	$b_3 c_3$	$b_3 \sigma_3$
. . .	. . .	. . .	. . .
$a_n b_n$	b_n^2	$b_n c_n$	$b_n \sigma_n$
Σab	Σb^2	Σbc	$\Sigma b\sigma$

From these tables we can easily pick out the required coefficients Σa^2, Σab, Σac, Σb^2 and Σbc.

The σ columns are used for checking.

E.g. (α) Each table checks itself by cross-addition.

(β) Each item in the σ columns of the 2nd and 3rd tables can be checked from the corresponding item of the σ

column in Table 1. Thus, $a_2\sigma_2$ in Table 2 = $a_2 \times \sigma_2$ in Table 1, &c.

Example 1. *Find the most probable values of x and y from the equations:*

$$-2x+3y = 4\cdot8$$
$$x+2y = 8\cdot1$$
$$\text{and} \quad 1\cdot7x+y = 6$$

by the method of least squares.

a	b	c	σ
−2	3	4·8	5·8
1	2	8·1	11·1
1·7	1	6	8·7
0·7	6	18·9	25·6

When it is more convenient to do so, as in the present instance, the c's which are on the right-hand sides of the equations may be dealt with as if they were on the left-hand sides, as in the standard equations, provided that Σac and Σbc in the resulting normal equations are put in the right-hand instead of the left-hand sides.

a^2	ab	ac	$a\sigma$
4	−6	−9·6	−11·6
1	2	8·1	11·1
2·89	1·7	10·2	14·79
7·89	−2·3	8·7	14·29

ab	b^2	bc	$b\sigma$
−6	9	14·4	17·4
2	4	16·2	22·2
1·7	1	6	8·7
−2·3	14	36·6	48·3

The normal equations are:

$$7\cdot89x-2\cdot3y = 8\cdot7$$
$$-2\cdot3x+14y = 36\cdot6.$$

Whence
$$\left.\begin{array}{l} x \doteq 1\cdot96 \\ y \doteq 2\cdot94 \end{array}\right\}$$

The three lines are shown in fig. 11. The most probable point (1·96, 2·94) is inside the little triangle made by the lines crossing.

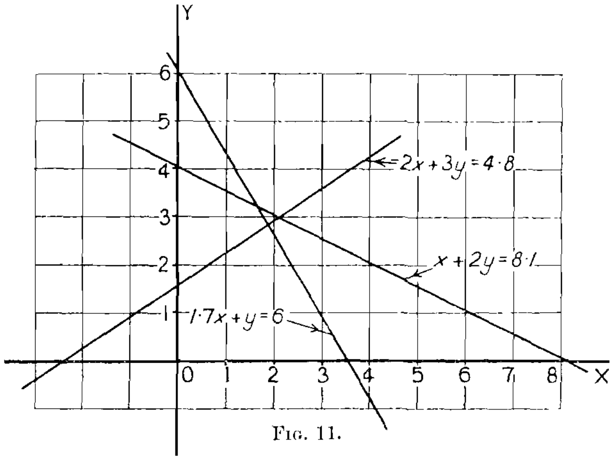


Fig. 11.

To find the equation to *the line of closest fit*, for a number of given points, assume the equation $y = ax + b$; * treat the a and b as variables and find their most probable values by the method of least squares.

Example 2. *Two variables, x and y, are supposed to be connected by an equation of the type $y = ax + b$. Find the most suitable values of a and b, from the following observed values of x and y.*

x	1	2	3	4	5
y	1·5	2·0	3·2	3·4	4·0

Comparing the equations $ax + b = y$, and $ax + by + c = 0$, and remembering that the a and b are to be treated as the variables, we see that the a, b and $-c$ of the latter equation are represented by x, 1 and y in the former. It is only necessary then to form tables of x, 1 and y on the same plan as those for a, b and $-c$. The effect of

* The equations $y = ax + b$ and $x = a'y + b'$ give rise to *two* lines of closest fit, intersecting at the C.G. of the system of points. (See footnote, p. 45.)

the sign of the c on the tabulation has already been mentioned. It should be noticed that the tables corresponding to the *first* and *third* of those used to find the most probable point, namely, $a\ b\ c\ \sigma$, and $ab\ b^2\ bc\ b\sigma$, will now be the *same* since the b is represented by unity.

I.

x	1	y	σ
1	1	1·5	3·5
2	1	2·0	5·0
3	1	3·2	7·2
4	1	3·4	8·4
5	1	4·0	10·0
15	5	14·1	34·1

II.

x^2	x	xy	$x\sigma$
1	1	1·5	3·5
4	2	4·0	10·0
9	3	9·6	21·6
16	4	13·6	33·6
25	5	20·0	50·0
55	15	48·7	118·7

The normal equations for a and b are:

$$15a + 5b = 14{\cdot}1 \quad . \quad . \quad . \quad \text{(i)}$$

$$55a + 15b = 48{\cdot}7 \quad . \quad . \quad . \quad \text{(ii)}$$

It may be pointed out here that the coefficient of a in equation (i) and that of b in equation (ii) are always identical because each is $\Sigma\,(x)$.

Solving for a and b, we have

$$a = 0{\cdot}64$$

$$b = 0{\cdot}90$$

The five points and the line $0{\cdot}64x + 0{\cdot}9 = y$ are shown in fig. 12.

When two measured quantities x and y are connected by equations of the types, $y = ax^n$ or $y = ab^x$, it is impossible to find the values of the constants by drawing the x-y graphs.

They can be found, however, by taking logarithms of both sides of the equations. The method is illustrated in the following examples.

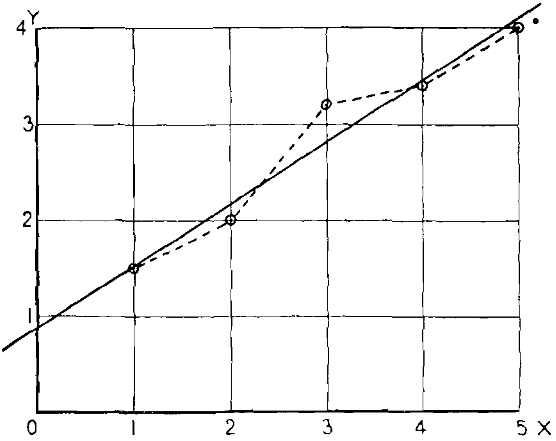


FIG. 12.

Example 1. *Find the equation connecting x and y from the following table :*

x	0	1	2	3	4	5
y	0	2.0	15.9	53.5	127	252

The equation is thought to be of the type, $y = ax^n$.

If $y = ax^n$ then

$$\log y = \log a + n \log x.$$

$$\text{or } Y = A + n X.$$

x	y	X	Y
0	0	$-\infty$	$-\infty$
1	2.0	0	0.30
2	15.9	0.30	1.20
3	53.5	0.48	1.73
4	127	0.60	2.10
5	252	0.70	2.40

The connexion $Y = 0{\cdot}30 + 3X$ is easily found, for $OA = 0{\cdot}30$ and $\dfrac{BC}{AB} = 3$ *(fig. 13)*.

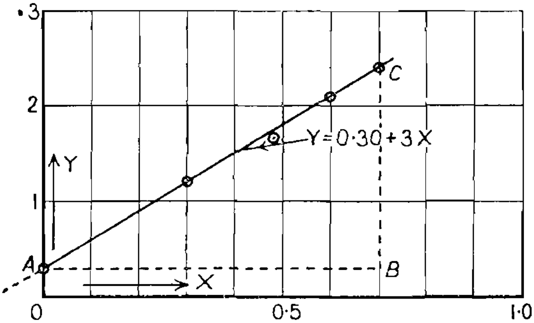


FIG. 13.

$$\therefore \log y = \log 2 + 3 \log x$$
$$\text{or} \quad y = 2x^3.$$

Example 2. *Find the equation connecting x and y, assuming it to be of the type $y = ab^x$, from the following set of corresponding values:*

x	0	1	2	3	4	5
y	10	19·8	41	79·3	162	318

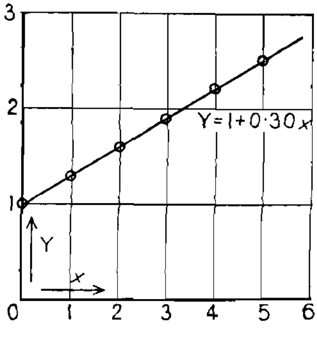


FIG. 14.

Taking logarithms we have:

$$\log y = \log a + x \log b,$$

or $$Y = A + Bx.$$

x	y	Y
0	10	1
1	19.8	1.30
2	41	1.61
3	79.3	1.90
4	162	2.21
5	318	2.50

From the graph (fig. 14) we obtain the equation $Y = 1 + 0{\cdot}30$.

$\therefore$ $a = 10$, and $b = 2$ and $y = 10 . 2^x$.

Exercises.

1. Find the most probable values of x and y, consistent with the equations:

$$x + y = 3,$$
$$-x + 2y = 6,$$
$$2x + y = 1{\cdot}5$$
$$-x + y = 3{\cdot}5.$$

Ans. $x = -0{\cdot}5$, $y = 2{\cdot}86$.

2. Find the equation of the line of closest fit for the following pairs of values of x and y, in the form $y = ax + b$.

x	0	1	2	3	4
y	40	80	150	200	240

Ans. $52x + 38 = y$.

3. Two quantities x and y should on theoretical grounds be connected by an equation of the form $y = ax^n$. Assuming this to be true, find the most probable values of a and n from the following table of associated values of x and y found experimentally.

x	0.5	1.5	2.5	5.0	10
y	3.4	7.0	12.8	29.8	68.2

Ans. $y = 4.3\,x^{1.2}$.

4. Two variables x and y are supposed to be connected by a relation of the form $y = ax + b$. Find the values of a and b which best fit the following observed values of x and y:

x	1	2	3	4	5
y	2.01	3.24	4.43	5.81	7.02

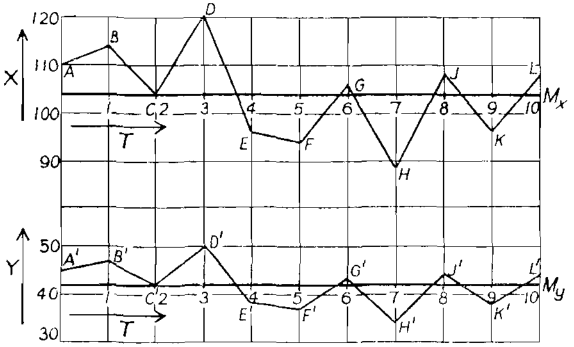
Ans. $a = 1{\cdot}26$.
,, $b = 0{\cdot}72$.

Prelim. Sci., Oxford (by permission).

VI

THE CORRELATION COEFFICIENT

FIGURES 15 and 16 show the variation of two quantities X and Y with respect to some other quantity T. The horizontal line along which T is measured is the mean value of each. Thus M_x the mean value of X is 104 and M_y the mean value of Y is 42.



FIGS. 15 and 16.

If $X - M_x = x$, and $Y - M_y = y$, we can read off the values of x and y (the variations from the means) by taking the lines M_x and M_y for zero-lines.

Fig. 17 shows M_x, M_y and the x-y graph, the lengths X and Y being measured from the top left-hand corner, as is usual with this type of graph. In order to make clear the method of drawing the graph, each point is marked with a small letter corresponding with the capital letters denoting the corresponding pair of points in the first two graphs.

The two quantities X and Y are in *perfect correlation*, because when any particular value of x is known, the corresponding value of y can be found with certainty. The variables x and y are, in fact, connected by the equations $x = 2y$ or $y = \frac{1}{2}x$.

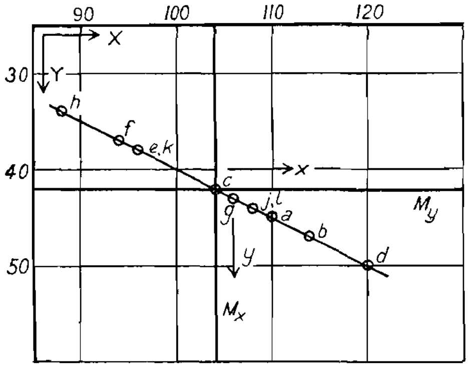
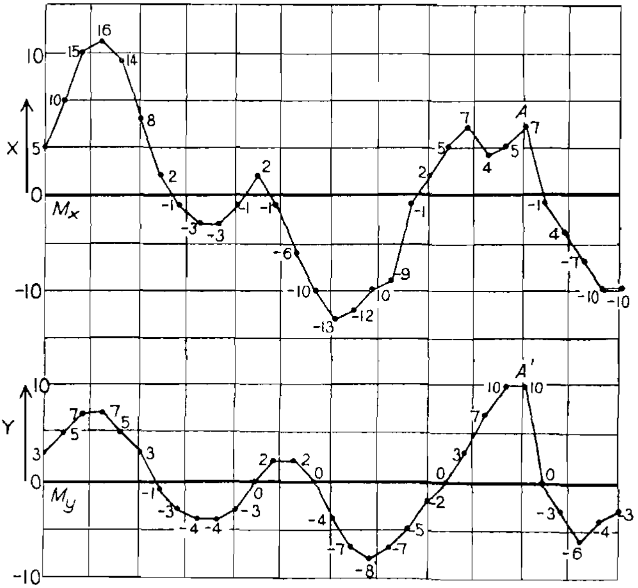


Fig. 17.



Figs. 18 and 19.

It often happens that two quantities, not being *perfectly* correlated, yet show a distinct connexion, high values of the one being associated with high values of the other and vice versa. One suspects that some underlying cause is affecting both quantities, and it is important for the scientist to find some measure of the correlation existing between them.

Figs. 18 and 19 show the deviations (x and y) of two quantities X and Y from their mean values M_x and M_y respectively. Although the curves are of quite different shapes, there seems to be a tendency for them both to fall and rise together. The quantities X and Y are certainly not *perfectly* correlated, for when x is 5, for example, y has 3 different values, 3, 0, and 10.

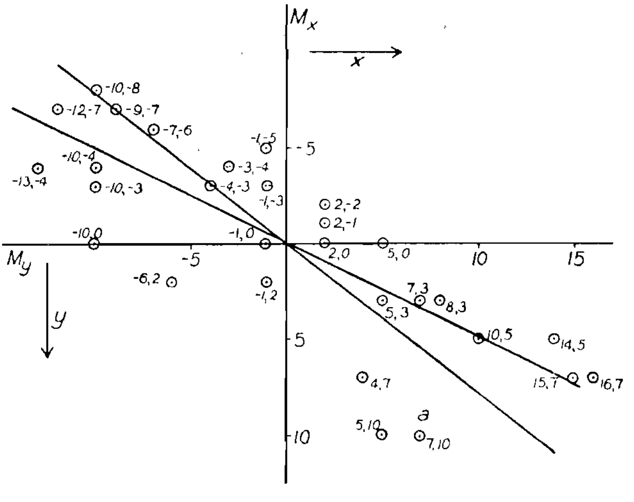


FIG. 20.

Draw two lines M_x, M_y at right-angles to represent the two means, as in fig. 20. Then plot in the points corresponding to each pair of points in figs. 18 and 19, exactly as was done in the case of perfect correlation (fig. 17). In fig. 20 each point is marked with its co-ordinates taking M_x and M_y for axes of x and y respectively, so as to make it clear which point corre-

sponds to each pair. For example, the points marked A, A' in figs. 18 and 19 are represented by a in fig. 20. The figure is called the *dot diagram*. Nothing is *obviously* discernible from it, except a general drift from NW. to SE., as it were.

When two points coincide they are represented by a double dot as in the case of $(-3, -4)$. The significance of the two cross-lines will be explained directly.

If an attempt is made to find the value of x corresponding to a particular value of y, we find that in some cases there is only one value of x, and in others there are several. For example, when $y = -4$, x has the four values, -3 (twice), -10 and -13, whereas when $y = -2$, x has the unique value, 2.

Draw a diagram, as in fig. 21, replacing multiple values of x (corresponding to one value of y) by their *mean* value, and leaving unique values *in statu quo*. For example, replace the 4 points, $(-3, -4)$, $(-3, -4)$, $(-10, -4)$, and $(-13, -4)$ by the one point $(-7\frac{1}{4}, -4)$. This point is marked A in fig. 21. Proceeding *downwards*, join the points and so obtain the broken-line graph. Now, with a piece of string, draw by eye the *line of closest fit*. This is the *regression line of x on y*.

In a similar way, by replacing multiple values of y (corresponding to one value of x) by their mean values, obtain the broken-line graph of fig. 22. The points are joined by proceeding from *left to right*. Draw, as before, the line of closest fit. This is the *regression line of y on x*.

In this example particular values of x and y have been associated for the sake of simplicity and to show clearly the connexion between the dot diagram and the two separate variation curves. In actual practice the points are usually plotted directly to form the dot diagram, and the means of the y's are obtained by dividing the diagram into columns by a series of vertical lines at regular intervals. The mean of the y's in one column is then plotted against the corresponding value of x. Similarly the means of the x's are obtained by dividing the diagram into rows by a series of horizontal lines

at regular intervals. The mean of the x's in one row is then plotted against the corresponding value of y. If a point lies exactly on a dividing line it is reckoned as $\frac{1}{2}$ in each of the columns or rows which lie on either side of it. If a point lies on the intersection of two dividing lines it is reckoned as $\frac{1}{4}$ in each of the four surrounding compartments. The procedure is illustrated in the worked example on p. 52.

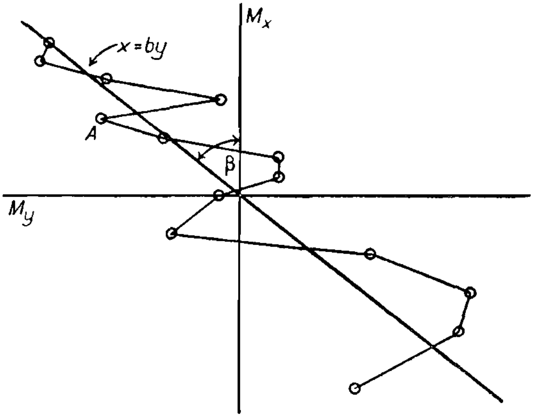


FIG. 21.

It will be noticed that both of the lines of closest fit pass through the origin. This is not accidental; it is necessarily so.* The line in fig. 21 is used to find values of x corresponding to particular values of y, and its equation is written $x = by$. Similarly the equation of the line in fig. 22 is written $y = ax$. If the lines $x = by$, and $y = ax$ make angles β and α with the y axis (M_x) and x-axis (M_y) respectively, $a = \tan \alpha$ and $b = \tan \beta$.

The two equations $x = by$ and $y = ax$ are called *regression equations*, of x on y, and y on x respectively.

* If $ax + b = y$ is a line of closest fit, then $a\Sigma(x) + nb = \Sigma(y)$ (see Table I, p. 37), and this is the condition that the point $\left(\frac{\Sigma(x)}{n}, \frac{\Sigma(y)}{n}\right)$, which is the C.G. of the system, lies on the line $ax + b = y$. (See footnote, p. 36.)

If there is any definite connexion between the x's and y's, the two regression lines must be situated either as in fig. 23 or as in fig. 24.

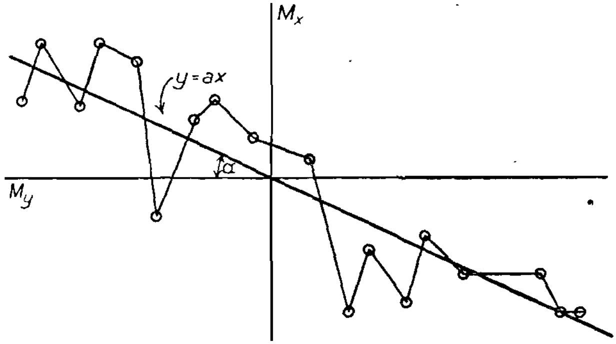


Fig. 22.

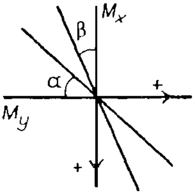


Fig. 23.

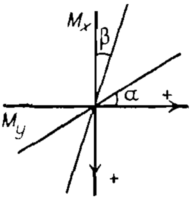


Fig. 24.

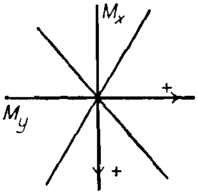


Fig. 25.

It would be impossible for the regression lines to lie in the positions of fig. 25, for in that case the same value of y would be associated with a positive *and a negative* value of x. In fig. 24 the correlation is *inverse*, that is, high values of x correspond with high values of y of *opposite sign*. This means *mathematically* that one of the curves X or Y of figs. 18 and 19 is *upside down*, and *physically* that the same underlying cause is, presumably, affecting both quantities, but in *opposite ways.* An instance of inverse correlation is that between the annual mean temperature of a place on the earth's surface and the

value of '*g*' there. As the latitude increases the former tends to decrease and the latter to increase.

Fig. 26 represents a regression line $x = by$ and the broken-line graph from which it was derived. Suppose a particular value of y is OA. Then the value of x actually corresponding to y is AC, but the value derived from the equation $x = by$ (where b is $\tan \beta$) is AB. Taking AC as the true value of x, CB will be the error (ϵ say).

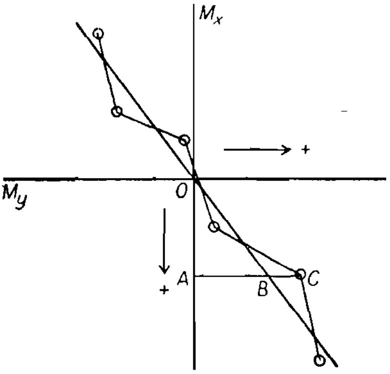


Fig. 26.

Then $x - by = -\epsilon$. Clearly the best position for the line of closest fit will be that for which $\Sigma(\epsilon)$ *irrespective of sign* is least. To eliminate the effect of the signs, we find the condition for which $\Sigma(\epsilon^2)$ is least. That is, we try to find the value of **b**, which will make $\Sigma(\epsilon^2)$ the least.

Differentiating $\Sigma(x - by)^2$ with *respect to b*, we have

$$\frac{\Delta\left[\Sigma(x-by)^2\right]}{\Delta(b)} = -2\Sigma xy + 2b\Sigma y^2.$$

If this function is zero, then $b = \dfrac{\Sigma(xy)}{\Sigma(y^2)}$,

Similarly $$a = \frac{\Sigma(xy)}{\Sigma(x^2)}.$$ 39

We will now proceed to find a correlation measure. We need only consider direct correlation, because, as is evident

from figs. 23 and 24, inverse correlation can be regarded mathematically as a mirror reflection of direct correlation, the line M_x being the mirror.

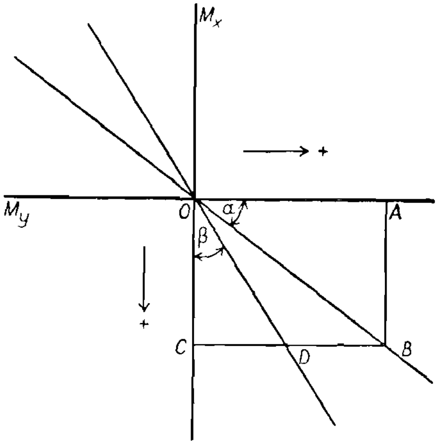


FIG. 27.

Suppose a particular value of x (OA) is chosen (fig. 27). The value y, found from the equation $y = ax$, is AB. But if this value of y is taken and used to find x from the equation $x = yb$, we get the value x' (CD).

$$\text{Now } \frac{x'}{x} = \frac{y \tan \beta}{y \cot \alpha} = \tan \alpha \tan \beta = ab = \frac{[\Sigma(xy)]^2}{\Sigma(x^2)\,\Sigma(y^2)}.$$

For perfect correlation $\alpha = 90° - \beta$ and $\tan \alpha \tan \beta = 1$.

That *unity* is the *upper limit* of $\tan \alpha \tan \beta$ is shown in Appendix, XXI. High values of the function indicate good correlation, and low values poor correlation.

It would appear from the preceding paragraph that $\tan \alpha \tan \beta$ would itself be an excellent correlation measurer. In fact, however, use is made of the *square root* of this function. The **coefficient of correlation, r,** is defined to be

$$\frac{\Sigma(xy)}{\sqrt{\Sigma(x^2)\,\Sigma(y^2)}}.$$

Since $r^2 < 1$ and > 0, r must lie between ± 1. By using r instead of r^2, we can discriminate between direct and inverse correlation. For good direct correlation $r \rightarrow +1$ and for good inverse correlation $r \rightarrow -1$.

The full working for the computation of a, b, and r for the dot-diagram of fig. 20 is as follows:

x	y	xy	x^2	y^2
5	3	15	25	9
10	5	50	100	25
15	7	105	225	49
16	7	112	256	49
14	5	70	196	25
8	3	24	64	9
2	−1	−2	4	1
−1	−3	3	1	9
−3	−4	12	9	16
−3	−4	12	9	16
−1	−3	3	1	9
2	0	0	4	0
−1	2	−2	1	4
−6	2	−12	36	4
−10	0	0	100	0
−13	−4	52	169	16
−12	−7	84	144	49
−10	−8	80	100	64
−9	−7	63	81	49
−1	−5	5	1	25
2	−2	−4	4	4
5	0	0	25	0
7	3	21	49	9
4	7	28	16	49
5	10	50	25	100
7	10	70	49	100
−1	0	0	1	0
−4	−3	12	16	9
−7	−6	42	49	36
−10	−4	40	100	16
−10	−3	30	100	9
		963	1960	760
		$\Sigma(xy)$	$\Sigma(x^2)$	$\Sigma(y^2)$

$$a = \frac{\Sigma(xy)}{\Sigma(x^2)} = \frac{963}{1960} = 0{\cdot}4913,$$

and $\alpha \triangleq 26° \; 10'$.

$$b = \frac{\Sigma(xy)}{\Sigma(y^2)} = \frac{963}{760} = 1{\cdot}267,$$

and $\beta \triangleq 51° \; 45'$.

These two values of a and b were used to draw the lines $x = by$ and $y = ax$ in fig. 20. It is interesting to notice that the values of β and α, as found in figs. 21 and 22 with a piece of string and a drawing-pin, have errors of less than $\frac{1}{2}°$ and about 1° respectively: *

$$r = \frac{\Sigma(xy)}{\sqrt{\Sigma(x^2)\,\Sigma(y^2)}}$$

$$= \frac{963}{\sqrt{1960 \times 760}}$$

$$\triangleq 0{\cdot}79.$$

* The errors are so small because the value of r is comparatively high. The method of drawing the regression lines by eye actually involves a fallacy, the effect of which, however, is usually negligible. The fallacy consists in the fact that each dot influences the eye to the same extent, whereas in fact, a dot which is the mean of 5 others (say) should carry 5 times as much weight as a unique dot. The point is discussed in the section on 'Weighted Observations' in ch. vii.

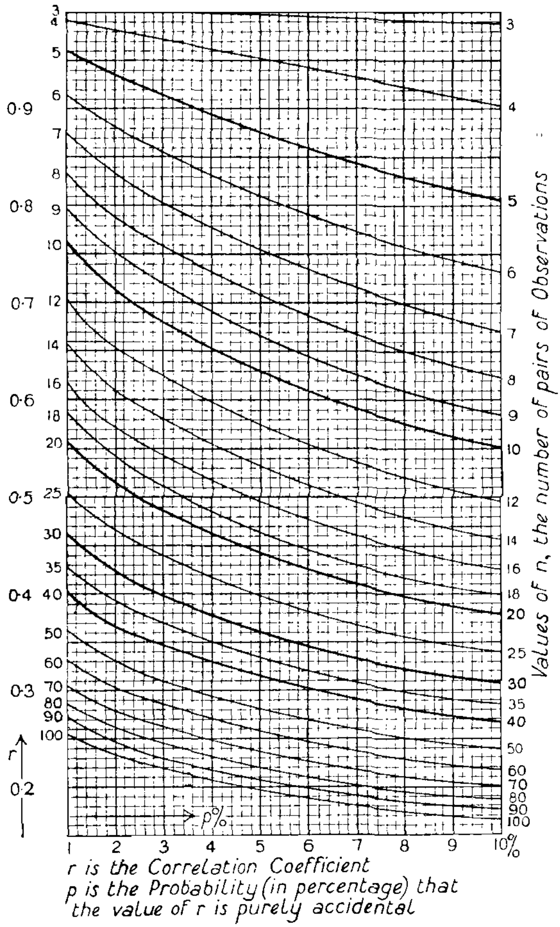


FIG. 28.

The significance of the Correlation coefficient.

When a scientist has obtained a value of *r* from a set of *n* pairs of observations, he naturally asks himself the question, 'Can I be sure that this value has a real significance, or have

I been unlucky enough to pick by sheer accident a set of dots giving me a spurious value of r, whose real value is zero, since there is no real correlation at all?' The subject has been investigated by Dr. R. A. Fisher, F.R.S., and the results published in his *Statistical Methods for Research Workers*,* to which the reader is referred for fuller information.

The following graph (fig. 28) is based on the figures given by Dr. Fisher. One can read from it at a glance the probability that any given value of r from a given number of pairs of observations, may be purely accidental. For example, suppose that from a dot-diagram of 20 dots, the correlation coefficient has been found to be 0·52. The probability of this being purely a chance result, and therefore of no real value, is about 1·8 per cent., or $\frac{18}{1000}$. The odds against the result being purely accidental are therefore about 982 : 18 or roughly 55 : 1.

The Probable Error of a correlation coefficient (ρ) is expressed by the formula $0{\cdot}6745\,\frac{1-\rho^2}{\sqrt{n}}$. As the true value ρ is not generally known one can only use the observed value r. When ρ is large the error due to using r instead of ρ may be too large for the result to be of any use. For small values of n, the formula itself fails, hence it should only be used for comparatively low values of r and high values of n (not less than 25).

The Contingency Table.

The computation of the values of $\Sigma(x^2)$, $\Sigma(y^2)$ and $\Sigma(xy)$ becomes very laborious when the number of observations is large. It is then better to proceed as follows: Rule the dot diagram into rows and columns. Estimate by eye the column and row in which M_x and M_y are most likely to lie; the errors are allowed for later.

Take the widths of the rows and columns, which need not be the same, for unit lengths. Measure all distances from the centre-lines of the M_x column and the M_y row, and assume

* For Dr. Fisher's exposition of the method on which his results are based, see *Biometrika* (1915), x. 507–21. and *Metron* (1924), iii. 329–32.

that each dot is at the centre-point of the compartment in which it lies, so that the co-ordinates of the points are all integral numbers. When a dot occurs on a dividing line, put a ring round it and replace it by 2 crosses, each of which is understood to represent $\frac{1}{2}$ a dot, one on each side of the dividing line.

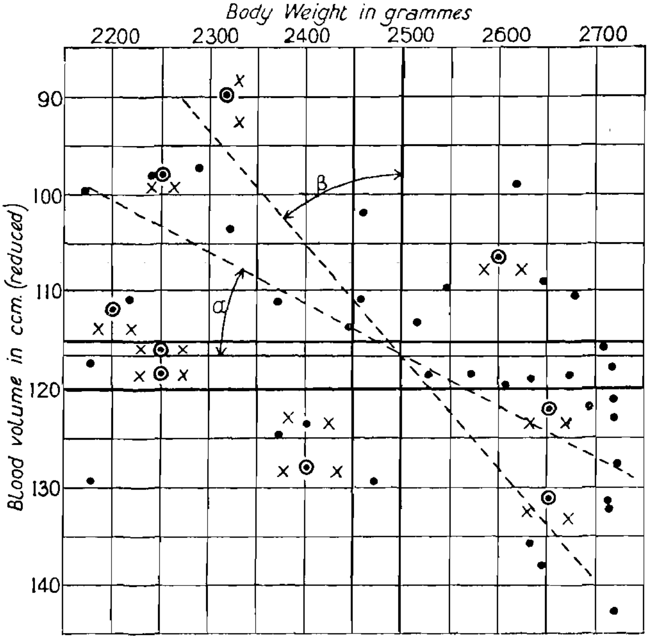


FIG. 29.

Further procedure is illustrated from a concrete example, taken from the *Collected Papers from the Department of Pathology in the University of Oxford*, by permission of Professor Dreyer, F.R.S. A series of 126 observations were made to find the correlation between the body-weight and blood-volume of certain warm-blooded animals.

The observations were tabulated according to body-weight. Fig. 29 is a dot-diagram of 44 of the observations for body-

weights ranging from 2,170 to 2,720 grammes. The column and row containing M_x and M_y respectively are enclosed in thick lines.

	-7	-6	-5	-4	-3	-2	-1	M_x	1	2	3	4	5	n_r	$n_r y$	$n_r y^2$
-6					½									½	-3	18
-5					½									½	-2½	12½
-4		1	1½	1½							1			5	-20	80
-3					1			1						2	-6	18
-2									1	½	1½			3	-6	12
-1		½	1½			1	1	1	1			1		7	-7	7
M_y		1	1	1					1	1	2	1	2	10	0	0
1						1½	½				½	1½	2	6	6	6
2		1				½	½	1					1	4	8	16
3											½	½	2	3	9	27
4											2			2	8	32
5													1	1	5	25
n_c		3½	4	2½	2	3	2	3	3	1½	7½	4	8	44	-8·5	253½
$n_c x$		-21	-20	-10	-6	-6	2	0	3	3	22½	16	40	+195		
$n_c x^2$		126	100	40	18	12	2	0	3	6	67½	64	200	638½		

FIG. 30.

Draw up a *Contingency table* on a framework of columns and rows exactly similar to those of the dot diagram. The mid-lines of the M_x column and the M_y row are taken for axes of x and y. In each compartment write down the number of dots which fall in it. The number of dots in a row is n_r and the number in a column n_c. Since the values of y are assumed to be the same in each row, the simplest way to evaluate $\Sigma(y)$ and $\Sigma(y^2)$ is to take the y-ordinates a row at a time, finding $n_r.y$ and $n_r.y^2$ for each row. It is obvious that $\Sigma(n_r.y)$ and $\Sigma(n_r y^2)$ equal $\Sigma(y)$ and $\Sigma(y^2)$ respectively. Similar remarks apply to the x-ordinates.

The Contingency table corresponding to the dot-diagram of fig. 29 is shown in fig. 30, together with the computation of $\Sigma(x)$, $\Sigma(y)$, $\Sigma(x^2)$ and $\Sigma(y^2)$.

Since $\Sigma(y) = -8{\cdot}5$, the true M_y line lies 0·19 of a unit (the breadth of a column) *above* the centre line of the M_x-column. Similarly, since $\Sigma(x) = +19{\cdot}5$, the true M_x line lies 0·44 of a unit to the *right* of the centre line of the M_x row, practically on the right-hand thick line. The corrections for these errors are given below.

To evaluate $\Sigma(xy)$, first write in each compartment which contains dots or half a dot, the value of xy for the centre of that compartment. In each compartment multiply the value of xy by the number of dots in it. Sum the values so obtained in 4 separate compartments as follows:

+117	−30
$-18\frac{1}{2}$	+117

whence $\Sigma(xy) = +185\frac{1}{2}$.

The values of $\Sigma(x^2)$, $\Sigma(y^2)$, and $\Sigma(xy)$ can be corrected for the errors in the assumed values of M_x and M_y.

Let d be the error in M_x, so that $M_x' = M_x + d$.

Then for every dot, $x = x' + d$ (for if d is $+^{ve}$ x' is clearly an *under*-estimate of x). Let n be the number of dots, then—

$$\begin{aligned}\Sigma(x^2) &= \Sigma(x'+d)^2 = \Sigma(x'^2) + 2d\,\Sigma(x') + nd^2\\ &= \Sigma(x'^2) + 2d\,[\Sigma(x) - nd] + nd^2\\ &= \Sigma(x'^2) - nd^2, \text{ since } \Sigma(x) = 0.\end{aligned}$$

Similarly if e is the error of M_y, $\Sigma(y^2) = \Sigma(y'^2) - ne^2$.

It can easily be shown in a similar way that $\Sigma(xy) = \Sigma(x'y') - nde$.

In the present example, $\Sigma(x^2) = 638{\cdot}5 - 44\,(0{\cdot}44)^2 \doteq 630$

$$\Sigma(y^2) = 253{\cdot}5 - 44\,(0{\cdot}19)^2 \doteq 252$$

$$\Sigma(xy) = 185{\cdot}5 - 44\,(0{\cdot}44)\,(-0{\cdot}19) \doteq 189.$$

$$\text{Hence} \quad r = \frac{189}{\sqrt{630 \times 252}} = 0{\cdot}474$$

$$\tan \alpha = \sqrt{\frac{189}{630}} = 28^\circ\ 43'$$

$$\tan \beta = \sqrt{\frac{189}{252}} = 40^\circ\ 54'.$$

The regression lines passing through the centre of gravity of the system of dots are shown by the broken lines in the diagram (fig. 29).

The value (0·474) of r from 44 observations is seen from the graph of fig. 28 to be materially significant since the odds are considerably more than 100 : 1 against the value being an accidental one.

VII

PARTIAL CORRELATION, WEIGHTED OBSERVATIONS, THE OBSERVATION ERROR, AND QUALITY CORRELATION

Suppose that a quantity Z is correlated with two other quantities X and Y, and that X and Y are correlated with one another. The correlation between X and Y will be partly influenced by the effect of Z. The problem is how to find the correlation between X and Y with the effect of Z eliminated.

Let r_{xy} be the correlation between X and Y and $r_{xy \cdot z}$ be the correlation between X and Y with the effect of Z eliminated, with a corresponding notation for the other pairs of variables. Then, it is shown in Appendix, XXV, that $r_{xy \cdot z}$ can be calculated from the formula,

$$r_{xy \cdot z} = \frac{r_{xy} - r_{xz} \cdot r_{yz}}{\sqrt{(1 - r_{xz}^2)(1 - r_{yz}^2)}}.$$

The method is here illustrated by consideration of a particular case. The age, weight, and height of ten athletes were taken, and in each case the heat produced in calories per 24 hours was measured. The experiment was actually on a much larger number, but a random selection of ten only has been made for the sake of simplicity.*

The particulars were as follows:

	Age in years.	*Weight in kilograms.*	*Height in centimetres.*	*Heat production in calories per 24 hours.*
	20	74·0	179	1914
	22	88·5	165	2017
	19	79·0	188	1944
	20	78·9	184	2126
	22	82·1	186	2034
	23	82·2	187	1978
	19	108·9	198	2559
	21	62·4	180	1816
	29	66·0	177	1695
	22	74·0	173	1908
	217	796·0	1817	19991
Means (approx.)	21·7	79·6	181·7	1999·0

* Quoted by permission from *A biometric study of basal metabolism in Man*, by J. A. Harris and F. A. Benedict (Carnegie Institute of Washington, U.S.A.).

Let a, w, h, and c be the deviations of the ages, weights, heights, and calories respectively from their means. Thus we have:

a	w	h	$c \div 10$
−1·7	−5·6	−2·7	−8·5
+0·3	+8·9	−16·7	+1·8
−2·7	−0·6	+6·3	−5·5
−1·7	−0·7	+2·3	+12·7
+0·3	+2·5	+4·3	+3·5
+1·3	+2·6	+5·3	−2·1
−2·7	+29·3	+16·3	+56·0
−0·7	−17·2	−1·7	−18·3
+7·3	−13·6	−4·7	−30·4
+0·3	−5·6	−8·7	−9·1

Leaving the ages out of account we shall proceed to find the correlation coefficients r_{cw}, r_{ch}, r_{wh} and thence to calculate $r_{cw \cdot h}$.

	w^2	h^2	$c^2 \div 100$	wh	$hc \div 10$	$cw \div 10$
	31	7	72	+15	+23	+48
	79	279	3	−149	−30	+16
	0	40	30	−4	−35	+3
	0	5	161	−2	+29	−9
	6	18	12	+11	+15	+9
	7	28	4	+14	−11	−5
	858	266	3136	+478	+913	+1641
	296	3	335	+29	+31	+315
	185	22	924	+64	+143	+413
	31	76	83	+49	+79	+51
Σ	1493	744	4760	505	1157	2482

Then
$$r_{cw} = \frac{24820}{\sqrt{476000 \times 1493}} = 0·931$$

$$r_{ch} = \frac{11570}{\sqrt{476000 \times 744}} = 0·615$$

and
$$r_{wh} = \frac{505}{\sqrt{1493 \times 744}} = 0·479.$$

Using the formula,

$$r_{cw \cdot h} = \frac{r_{cw} - r_{ch} \times r_{wh}}{\sqrt{(1-r_{ch}^2)(1-r_{wh}^2)}}$$

we have:

$$r_{cw \cdot h} = \frac{0{\cdot}931 - 0{\cdot}615 \times 0{\cdot}479}{\sqrt{(1-0{\cdot}615^2)(1-0{\cdot}479^2)}} = \frac{0{\cdot}637}{\sqrt{0{\cdot}622 \times 0{\cdot}771}} = 0{\cdot}920.$$

Having found the coefficient of the correlation between the heat production per 24 hours and the weight, with the effect of the height eliminated, it is possible to go a step farther and to eliminate the effect of the age.

The coefficient of correlation between the heat and the weight with the effects of both height and age eliminated is designated by the symbol $r_{cw \cdot ah}$.

In general, if x, y, u, and v are four variables, the coefficient of correlation between x and y with the effects of both u and v eliminated as far as possible, can be calculated from the formula,

$$r_{xy \cdot uv} = \frac{r_{xy \cdot v} - r_{ux \cdot v} \times r_{uy \cdot v}}{\sqrt{(1-r_{ux \cdot v}^2)(1-r_{uy \cdot v}^2)}}. \qquad 43$$

The graph on p. 50 can be used for partial correlation coefficients, provided that n is replaced by $(n-1)$ or $(n-2)$ according as one or two variables have been eliminated. Thus for a value of $r_{xy.z}$ from 20 sets of observations, the curve corresponding to $n = 19$ should be used.

Weighted and grouped observations.

Observations differ in respect to reliability, and it is customary to assign to different observations numbers to represent their relative reliabilities. These numbers are the *weights* of the observations. An observation weighted 7, for example, is regarded as equivalent to 7 observations of unit weight or reliability.

When a set of observations is *grouped* and the centre of gravity of each group is represented by one dot on the dot-

diagram, it is clear that the weight of each dot is the number of observations from which it was derived, if we neglect the 'scatter' of the group.

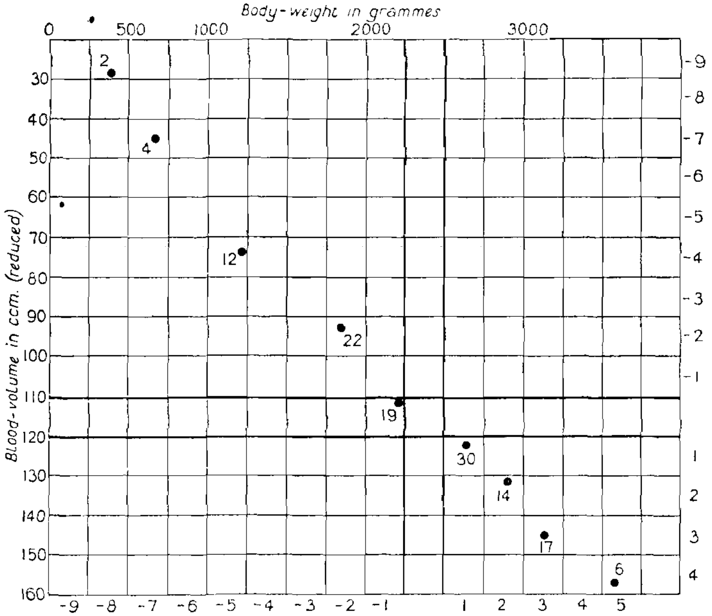


Fig. 31.

The set of 126 observations on body-weight and blood-volume referred to on p. 52 was arranged in 9 groups, and the mean values of the variables determined for each group. The results are displayed on the dot-diagram of fig. 31. The number beside each dot is its weight, the one marked 22, for example, was derived from a group of 22 observations.

The computation of the correlation coefficient is as follows:—

n_r	$n_r y$	$n_r y^2$	n_c	$n_c x$	$n_c x^2$
2	−18	162	2	−16	128
4	−28	196	4	−28	196
12	−48	192	12	−60	300
22	−44	88	22	−44	88
19	0	0	19	−19	19
30	30	30	30	30	30
14	28	56	14	28	56
17	51	153	17	51	153
6	24	96	6	30	150
	−5	973		−28	1120

$\Sigma\ xy)$

+668	0
0	+359

$\Sigma\,(xy) = 1027$ (uncorrected)

Corrections: $M_x = -\frac{48}{126} = -0{\cdot}22,\ M_y = -\frac{5}{126} = -0{\cdot}04.$

$$\Sigma\,(x^2) = 1120 - 126\,(0{\cdot}22)^2 = 1114$$

$$\Sigma\,(y^2) = 973 - 126\,(0{\cdot}04)^2 = 973$$

$$\Sigma\,(xy) = 1027 - 126\,(0{\cdot}22)\,(0{\cdot}04) = 1026$$

$$r = \frac{1026}{\sqrt{973 \times 1114}} = 0{\cdot}99.$$

It should be observed that the effect of grouping the observations is to make the value of r meretriciously high by eliminating the 'scatter' of the rows and columns. On the other hand, this treatment of the data may discover a real relationship which might otherwise be disguised or hidden by the effect of the scatter.

The Observation Error.

In any dot-diagram errors of observation are bound to reduce the observed correlation coefficient, for if a quantity cannot be measured accurately its variations are bound to be partially hidden under the errors of observation.

Suppose that x is a true value and x' a measured value where $x' = x + a$, and a is the error of observation. Similarly, suppose that $y' = y + b$.

It is shown in Appendix XXII that,

$$r = r'(1 + E^2)$$

where

$$E^2 = \tfrac{1}{2}\left\{\left(\frac{\sigma_a}{\sigma_x}\right)^2 + \left(\frac{\sigma_b}{\sigma_y}\right)^2\right\}.$$

In general $\frac{\sigma_a}{\sigma_x}$ and $\frac{\sigma_b}{\sigma_y}$ are about equal. Assuming that the equality is perfect, $E = \frac{\sigma_a}{\sigma_x} = \frac{\sigma_b}{\sigma_y}$.

Since $r = r'(1 + E^2)$.

$\therefore \quad E^2 = \frac{r - r'}{r'}$.

Now the true value of r is not known, so it is conventional to take it at its maximum value, unity, and to say that the

$$\textbf{Observation Error } (E) = \sqrt{\frac{1-r'}{r'}} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad \mathbf{4}$$

If the correlation is really perfect, and the difference $1 - r'$ is entirely due to errors of observation, $r'E^2$ is a measure of the effect of these errors, but if, as is usually the case, the real value of r is not known, E merely sets a maximum limit to the errors of observation. Thus if $r' = 0{\cdot}9$, then $E^2 = \frac{1}{9}$ and $E = \frac{1}{3}$ numerically. It follows that neither $\frac{\sigma_a}{\sigma_x}$ nor $\frac{\sigma_b}{\sigma_y}$ can exceed $\frac{1}{3}$, in other words that the standard deviation of errors of observation of x and y cannot exceed $\frac{1}{3}$ of the standard deviation of x and y respectively. This is only a limiting value; actually $\frac{\sigma_a}{\sigma_x}$ and $\frac{\sigma_b}{\sigma_y}$ may each be much less than $\frac{1}{3}$.

The significance of E may perhaps be more clearly realized from an actual example.

Suppose the two regression lines obtained from a series of observations are OP and OQ (fig. 32) where $\alpha' = M_y\widehat{O}P = 35°$ and $\beta' = M_x\widehat{O}Q = 27° \, 13'$.

Then $\qquad r' = \sqrt{\tan\alpha' \tan\beta'}$

$\qquad\qquad = 0{\cdot}6.$

On the assumption that the true correlation was perfect and represented by 2 coincident lines like OT, the maximum limit of E would be $\sqrt{\frac{1 - 0{\cdot}6}{0{\cdot}6}} = 0{\cdot}82.$

But practically perfect observations might show that the true correlation would be represented by the regression lines OR and OS, where $M_y \widehat{O} R = a = 39^\circ\, 19'$ and $M_x \widehat{O} Q = \beta = 38^\circ$.

Then $$r = \sqrt{\tan a \tan \beta} =$$

and $$E = \sqrt{\frac{r-r'}{r'}} \left(\text{instead of } \sqrt{\frac{1-r'}{r'}}\right)$$

and $$E = \sqrt{\frac{0 \cdot 2}{0 \cdot 6}} = 0 \cdot 58.$$

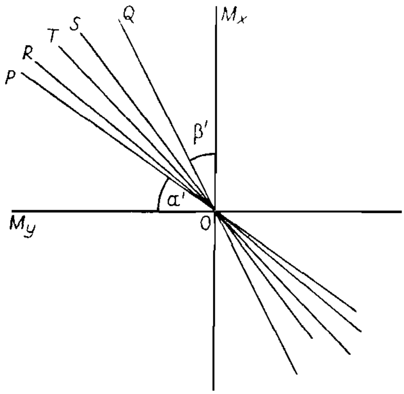


FIG. 32.

Quality Correlation.

This kind of correlation will perhaps be most easily understood from a concrete example. Let us suppose that there have been 50 cases of a rare disease among cattle in different parts of the world, and with different breeds of cattle. It is required to find out if possible what correlation exists between the breeds and the localities as regards liability to the disease. The countries in which the diseases are supposed to have occurred are The Argentine, Canada, Brazil, and Australia, indicated by the letters a, β, γ, and δ. The breeds were A, B, C, D, and E. The classification was as follows:—

	A	B	C	D	E	Total
α	6	4	3	0	1	14
β	4	3	3	1	0	11
γ	2	0	1	5	2	10
δ	1	1	3	7	3	15
Total	13	8	10	13	6	50

The last row of the table gives the distribution of the victims of the disease according to the breeds. If the localities and the breeds were entirely independent, each locality would have the breeds distributed in the same way. For example, in locality β, the number of cases occurring with breed *D* would be $\frac{13}{50}$ of 11, approximately 2·9, instead of only 1. The difference, 1·9, is supposed to be due to the influence of the locality. It is clear, therefore, that the correlation between the breeds and the localities must be some function of the differences calculated in this way.

Let x be the actual number in a compartment, and x' the number that should be in it if the localities and breeds were independent.

Then
$$\sum \frac{(x-x')^2}{x'} = n\phi^2$$

where ϕ is the *total mean square contingency*, and n is the number of observations,

and
$$c = \sqrt{\frac{\phi^2}{1+\phi^2}}$$

where c is the *coefficient of mean square contingency*. It is clear from the definitions that c must lie between 0 and 1 and that small values of c correspond to small correlations.

The calculation of the quantity $\frac{(x-x')^2}{x'}$ for locality γ and breed *D* is as follows:

$$x = 5,\ x' = \frac{13}{50} \times 10 = 2{\cdot}6$$

$$\therefore\ (x-x') = 2{\cdot}4$$

and
$$\frac{(x-x')^2}{x'} = \frac{5{\cdot}8}{2{\cdot}6} = 2{\cdot}2.$$

For the whole tabulation,

$$\sum \frac{(x-x')^2}{x'} = 22{\cdot}2 = 50\phi^2$$

$$\therefore \quad \phi^2 = \frac{22{\cdot}2}{50} = 0{\cdot}444$$

and

$$c = \sqrt{\frac{0{\cdot}444}{1{\cdot}444}} = 0{\cdot}55.$$

VIII

THE CORRELATION RATIOS

Failure of the Correlation Coefficient.

FIG. 33 is a dot-diagram of two quantities, X and Y. The connexion between them is obviously parabolic and there are no real linear regression lines. Yet, because there are only a few positive values of xy and a large number of negative values, $\Sigma(xy)$ is large numerically, and in consequence the value of r_{xy} is comparatively high and negative. The regression lines obtained by calculation lie as shown in the figure, but they do not correspond to any real linear regression at all.

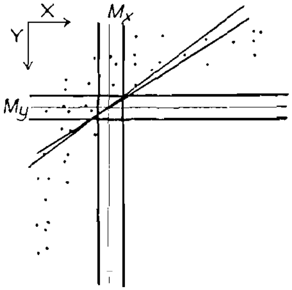


FIG. 33.

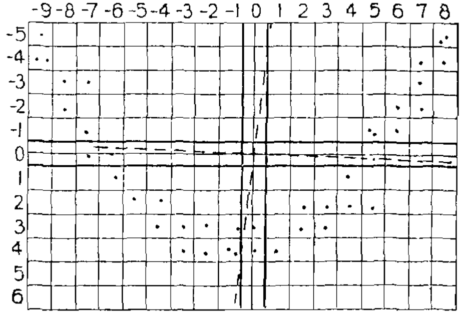


FIG. 34.

Fig. 34 shows that X and Y are definitely related in some way; the regression *curve* of y from x would approximate to the equation $y = K \cos x$. Yet because there are about equal

small. Actually $\Sigma(xy) = -46$, $\Sigma(x^2) = 1245$ and $\Sigma(y^2) = 358$, giving the values $r_{xy} = 0{\cdot}069$, $\alpha = 2°$ and $\beta = 7\frac{1}{2}°$. These calculated regression lines are shown in the figure, but they are clearly quite worthless.

The Correlation Ratios.

When the curves of regression are not linear it is better to use the *Correlation Ratios*. Corresponding to the two coefficients a and b of the equations $y = ax$ and $x = by$ there are two correlation ratios of 'y from x' (η_{yx}) and 'x from y' (η_{xy}).

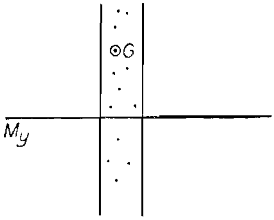


FIG. 35.

Consider a particular column of a dot-diagram and regard the dots as small masses each of unit weight (fig. 35). Suppose there are n_c of them in the column.

Let the distances of the dots from M_y be $y_1, y_2, y_3 \cdots y_{n_c}$ and the distance from G to M_y be m_y.

Let
$$y_1 = m_y + d_1$$
$$y_2 = m_y + d_2$$
$$\cdot \quad \cdot \quad \cdot \quad \cdot$$
$$y_{nc} = m_y + d_{nc}.$$

Now since the Moment of Inertia of the whole column of dots each of unit mass about M_y is equal to the Moment of Inertia of the dots about G plus the whole mass $\times m_y^2$,

$$\therefore \quad \Sigma(y^2) = \Sigma(d^2) + n_c m_y^2$$

or
$$\Sigma(d^2) = \Sigma(y^2) - n_c m_y^2.$$

Then for the *whole system*,

$$\Sigma(d^2) = \Sigma(y^2) - \Sigma(n_c m_y^2)$$
$$= \Sigma(y^2)\left[1 - \frac{\Sigma(n_c m_y^2)}{\Sigma(y^2)}\right]$$
$$= \Sigma(y^2)\left[1 - \eta_{yx}^2\right]$$

where
$$\eta_{yx} = \sqrt{\frac{\Sigma(n_c \cdot m_y^2)}{\Sigma(y^2)}}.$$

Similarly $$\eta_{xy} = \sqrt{\frac{\Sigma\,(n_r m_r{}^2)}{\Sigma\,(x^2)}},$$

where η_{xy} is the Correlation Ratio of 'x from y'.

Since $\Sigma(y^2)$ is the same as $\Sigma(n_r y^2)$,* we may take

$$\eta_{yx} \text{ as } \sqrt{\frac{\Sigma\,(n_c m_y{}^2)}{\Sigma\,(n_r \cdot y^2)}}$$

and $$\eta_{xy} \text{ as } \sqrt{\frac{\Sigma\,(n_r m_r{}^2)}{\Sigma\,(n_c x^2)}}.$$

These forms are the more suitable for computation.

Since $$\Sigma\,(d^2) = \Sigma\,(y^2)\,(1-\eta_{yx}{}^2)$$

$$1-\eta_{yx}{}^2 = \frac{\Sigma\,(d^2)}{\Sigma(y^2)},$$ which must be positive;

$\therefore\ \eta_{yx}$ cannot exceed 1.

When $\eta_{yx} = 0$, $\Sigma(d^2)$ has its maximum value $\Sigma(y^2)$, and $\dfrac{\Sigma(n_c m_y{}^2)}{\Sigma(n_r y^2)} = 0$. This latter relation is only true when each separate m_y vanishes; that is, when the Centres of Gravity of all the columns lie on the line M_y.

When η_{yx} has its maximum value, unity, $\Sigma(d^2)$ has its minimum value, zero, which will be the case when all the dots in each row are concentrated at its Centre of Gravity.

The connexion between r and η_{yx} is shown in the equation

$$\Sigma\,(n_c l^2) = \Sigma(y^2)(\eta_{yx}{}^2 - r^2)$$

where l in each column is the vertical distance between the Centre of Gravity of all the dots and the regression line, $y = ax$.

The formula is proved in Appendix, XXVI.

$(\eta_{yx}{}^2 - r^2)$ is a measure of the *deviation from linearity* of the curve of regression.

In general $\eta_{yx} - r^2 > 0$;

$$\therefore\ \eta_{yx} > r.$$

When $\eta_{yx} = r$, $\Sigma(n_c l^2) = 0$, and this can only happen when the Centres of Gravity of all the columns lie on the regression line.

* See Contingency Table, p. 53.

We shall illustrate the computation and interpretation of the Correlation Ratios by reference to the dot diagram of fig. 34.

To calculate η_{yx}, rule four columns for computation, headed n_c, m_y, m_y^2, and $n_c m_y^2$. For the sake of example, take the 3rd column from the left. There are three dots in it, so $n_c = 3$. The Centre of Gravity of the dots is $\frac{-3-1+0}{3} = -1{\cdot}3$, to one decimal place, so m_y is $-1{\cdot}3$, whence $m_y^2 = 1{\cdot}7$ and $n_c m_y^2 = 5{\cdot}1$.

The reader who wishes to work through the computation can check his working from the following:—

η_{yx}

n_c	m_y	m_y^2	$n_c\,m_y^2$
3	−4·3	18·5	55·5
2	−2·5	6·3	12·6
3	−1·3	1·7	5·1
2	0·5	0	0
1	2·0	4	4
2	2·5	6·2	12·4
2	3·5	12·3	24·6
2	3·5	12·2	24·4
2	4·0	16·0	32·0
2	3·5	12·3	24·6
1	4·0	16·0	16·0
2	2·5	6·2	12·4
2	2·5	6·3	12·6
2	1·5	2·2	4·4
4	0	0	0
2	−1·5	2·3	4·6
3	−3·0	9·0	27·0
3	−4·7	22·1	66·3

$\Sigma (n_c m_y^2) = 338{\cdot}5$

η_{xy}

n_r	m_x	m_x^2	$n_r\,m_x^2$
3	+2·3	5·3	15·9
4	−0·8	0·6	2·4
3	−2·7	7·3	21·9
3	−1·7	2·9	8·7
4	+2·2	4·8	19·2
3	−2·7	7·3	21·9
2	−1·0	1·0	2·0
6	+0·8	0·6	3·6
6	−0·7	0·5	3·0
6	−1·0	1·0	6·0

$\Sigma (n_r m_x^2) = 104{\cdot}6$

$\Sigma (y^2)$ and $\Sigma (x^2)$ computed as on p. 53 are 358 and 1245 respectively.

There are two points in connexion with the computation which should be mentioned before proceeding further. Nine of the values of m_y^2 and several of those of m_x have 5 as the only figure discarded. The values recorded are therefore, in order to allow as far as possible for this error, alternately high and low. For example 6·25, 6·25, 12·25, and 12·25 are recorded as 6·3, 6·2, 12·3, and 12·2.

Again, M_x and M_y the means of X and Y were assumed to be 10 and 6 respectively. Actually the means are 9·875 and 6·15, so that the errors are +0·125 and −0·15. These errors can

be allowed for as explained on page 54 under the heading 'Contingency Table'. As the differences due to these errors are small (in each case less than 1) they have been neglected.

$$\eta_{yx} = \sqrt{\frac{\Sigma(n_c m_y^2)}{\Sigma(n_r y^2)}} = \sqrt{\frac{338 \cdot 5}{358}} = 0 \cdot 97$$

$$\eta_{xy} = \sqrt{\frac{\Sigma(n_r m_x^2)}{\Sigma(n_c x^2)}} = \sqrt{\frac{104 \cdot 6}{1245}} = 0 \cdot 29.$$

It is not difficult to interpret these results. Fig. 36 shows the positions of the Centre of Gravity of the columns, based on the column headed m_y. (Note that each m_y is measured from the mid-line of the M_y row.) The regression curve of y from x has been drawn by eye.

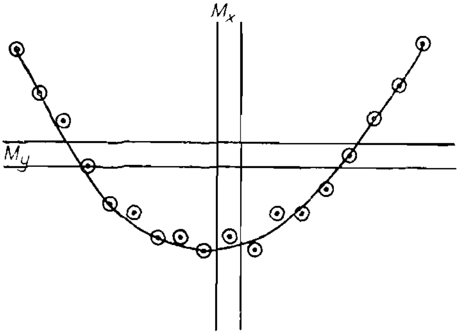


FIG. 36.

The value $\eta_{yx} = 0 \cdot 97$ is high because the centres of gravity lie so closely clustered about the curve of regression.

From the fact that $\eta_{xy} = 0 \cdot 29$ is low, we should expect that the centres of gravity of the rows would lie near the line M_x, and that is exactly what does happen, for the line M_x is approximately an axis of symmetry.

The Probable Error ($P\eta$) of a Correlation Ratio is given by the formula, $P\eta = 0 \cdot 6745 \frac{1 - \eta^2}{\sqrt{n}}$.

IX

THE ERROR OF THE LAST FIGURE, THE ERROR OF THE INTERVAL, ERRORS OF FUNCTIONS

The Error of the Last Figure.

SUPPOSE some quantity such as a logarithm is known to a comparatively large number of decimal places and that it is expressed to 4 *significant figures*. What kind of error is introduced by cutting off the last figure and expressing it to the 3rd significant figure? It is the same kind of error as that of a measurement recorded by an instrument graduated to read to 1 decimal place compared with that of another one reading to 2 decimal places. In order to consider this error apart from others it is necessary to assume provisionally that the reading of the more accurate instrument is actually, as it is relatively, 'correct'. This kind of error may be called the 4
Error of the last figure.

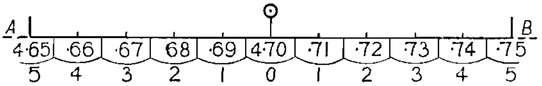


FIG. 37.

We have, let us say, a large number of distinguishable points evenly distributed along a straight line AB, on which a graduated transparent scale such as that of fig. 37 is placed. If the graduation were a tenth of a unit apart all the points between 4·65 and 4·75 would be classed as 4·7, but assuming that the graduation marks took up no space, and that they were a hundredth of a unit apart, each point could be discriminated as lying in one or other of the smaller intervals.

It will be noticed that whether we consider the positive and negative errors, grouped in this way, separately or together, the space corresponding to an error of 5 (actually 0·05) is only half the space corresponding to an error of 1, 2, 3, or 4. The same applies to an error of 0.

We should expect the curve of these 'errors of the last figure' considered *numerically* to take the form of that in fig. 38.

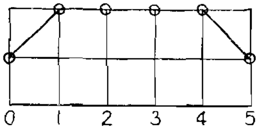


FIG. 38.

In order to supply a concrete instance of this kind of error an experiment was arranged on the following plan. A series of 263 numbers (of imports and exports) were taken at random from *Whitaker's Almanack*. These numbers expressed to two degrees of accuracy differing by one significant figure were written down and compared. In cases where the more accurate number ended in a 5, reference was made to the number from which it was derived to settle whether the 5 should be considered more or less than half of a unit's digit to the left of it.

The errors of the last figure were found to be distributed as follows :

Numerical Error.	*Positive.*	*Negative.*	*Positive and Negative.*
0	13	16	29
1	28	23	51
2	29	24	53
3	37	20	57
4	33	15	48
5	13	12	25
			263

Fig. 39 is a graph showing the connexion between the sums of the positive and negative errors and their frequencies. Replacing the sums of the 0 and 5 errors (29 and 25 respectively) by their average value, 27, and the sums of the errors of 1, 2, 3, 4 by their average value, 52, we obtain the dotted line graph, which is a very close approximation to the ideal curve of fig. 38.

Considering these errors as those of the less accurate instrument compared with the more accurate one, a glance at the

figure is sufficient to show that the *probable error of the last figure* is equal to *one-quarter of the interval of the less accurate instrument, or one-quarter of a last significant unit of the less* 49
accurate number.

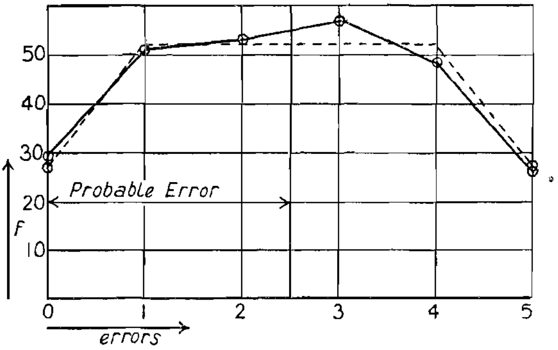


FIG. 39.

These errors of the last figure originate solely from our system of numeration. Actually, real errors would fall into an infinite number of compartments instead of five on each side. Suppose there were n compartments instead of 5. Then the curve of distribution would slope downwards in just the same way in the two end compartments, namely that from 0 to 1 and that from $(n-1)$ to n. It follows that if n were infinite the *curve of distribution* would be *a horizontal straight line* from half the interval on the left to half the interval on the right.

The Error of the Interval.

We shall call the error caused solely by the graduation of an instrument the *error of the interval.* It is clear from the above that *the probable error of the instrument is the same as the probable error of the last figure, namely one-quarter of the interval.*
It is just as likely to lie between a quarter interval on the left 50
and a quarter interval on the right, as it is to lie beyond those limits. It should be noted that in laboratory practice the

working interval is not always the same as the interval of the scale. In reading for example a burette as 34·27 c.cm. the operator cannot state with certainty that the quantity measured lies between 34·265 and 34·275 c.cm., but only that it lies between (say) 34·26 c.cm. and 34·28 c.cm. The working interval in this case would be 0·02 c.cm. and not 0·01 c.cm. as the scale indicates.

The important question arises: when 4 or 5 measured quantities are added or subtracted together, how are the errors of the instrument going to affect the result? For example, in the sum of 5 such quantities the error of the interval *only*, apart from all others, *may* be as much as $2\frac{1}{2}\times$ interval of the instrument. Is this likely to be the case? Is it sufficiently probable to be taken into consideration?

We may assume that when the number of errors between two successive graduation marks is infinite they occur at regular intervals, indefinitely near to one another. Suppose that the number of intervals is r, where r is a large number. Consider the function

$$[x^{-r}(1+x+x^2+\ \dots\ x^{2r})]^n.$$

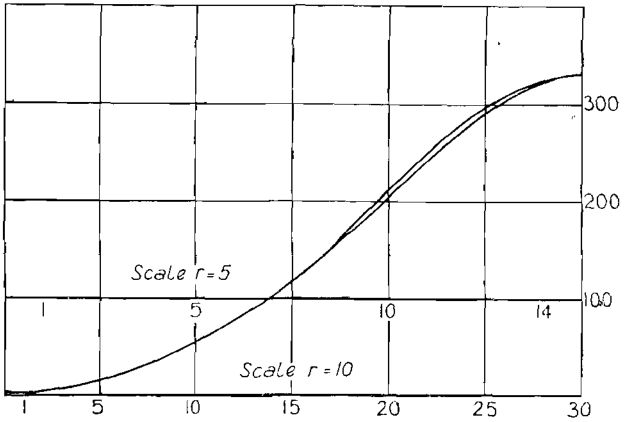
In the expanded expression the indices represent the errors that may occur when n measurements are added or subtracted and the coefficients are proportional to the frequencies, each to each. Perhaps the reason for this will be clearer if the function is written $(x^{-r}+x^{-r+1}+\ \dots\ 1+x+\ \dots\ x^{r-1}+x^r)^n$. Each error of the sum of n may be made by a choice of n of the possible errors $-r$, $-r+1$, ... 0, 1, 2, ... r.

The labour of expanding the expression when r is large is prohibitive. In fig. 40 are shown two curves which for the greater part are indistinguishable. The upper one is the error-frequency curve deduced from the expansion to the first 30 terms of

$$(1+x+x^2+\ \dots\ x^{20})^3,\ \text{where}\ r=10$$

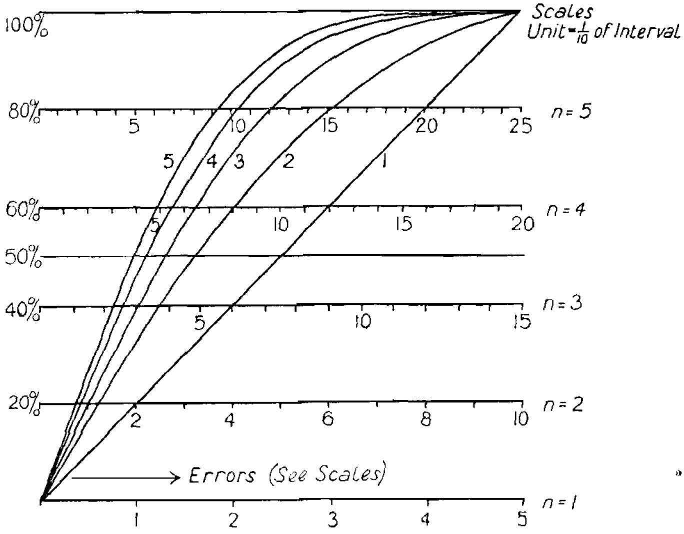
and the lower one that from the expansion to the first 15 terms of

$$(1+x+x^2+\ \dots\ x^{10})^3,\ \text{where}\ r=5.$$



Scale r=5
Scale r=10
300
200
100
1
5
10
14
1
5
10
15
20
25
30

FIG. 40.



Scales
Unit = 1/10 of Interval
100%
80%
60%
50%
40%
20%
n=5
n=4
n=3
n=2
n=1
5
4
3
2
1
Errors (See Scales)

FIG. 41.

As the difference is so little although r is twice as big in one case as it is in the other, it may be safely concluded that a study of the expansion of $(1+x+x^2+\ldots x^{10})^n$ will be sufficient to enable us to draw practical conclusions with sufficient accuracy for our purpose.

Fig. 41 shows the percentage-error curves (see p. 21) for the addition or subtraction of 2, 3, 4, and 5 measured quantities, reduced to a common vertical scale. The errors referred to are, of course, those resulting from errors of the interval only, apart from all others. They are deduced from the expansions of $(1+x+x^2\ldots x^{10})^n$ when $n=2, 3, 4$, and 5. This tacitly assumes that there are five errors on the right, five on the left, and one error of 0, as is at once apparent when the expression to be expanded is written in the form $x^5(x^{-5}+x^{-4}+\ldots x^0+\ldots x^4+x^5)$. This corresponds to the case where the errors are tenths of the interval of the instrument.

The figures from which the curves are drawn are given in Appendix, XXIV.

From the errors derived from the expansions the corresponding Probable Errors can be deduced empirically, and the results are exhibited in the following table:

$n=$	1	2	3	4	5	6
$P=$ (in terms of the interval	0.25	0·32	0.39	0.44	0·49	0·54
(in terms of the maximum error)	0·25	0.16	0·13	0.11	0.10	0·09

(See Appendix, XXIV.)

The chances that the maximum errors will occur are extremely small. For example, when $n=5$ the odds against the combined error exceeding $\frac{4}{5}$ of the maximum error are 425 : 1, and the odds against it exceeding 90 per cent. of the maximum are 3096 : 1.

It may be observed that the foregoing deductions apply with equal force to the use of tables of limited accuracy, such as 4-figure logarithm tables. Were it not for the fact that in

practice the natural law of chance tends to minimize the probability of the maximum error they would be of much less use than they are.

Since all Arithmetical processes can be analysed into steps in Addition it follows that Errors due to the limits of Arithmetical accuracy, that is *Arithmetical Errors*, must be of the *Interval* type. Moreover, if every step of the working is carried to the same number of significant figures as are given in the *least* accurate of the data the relative error of the result cannot exceed the relative error of that datum (**52**). It is usual to carry the working of each step to one *more* significant figure than is given in the least accurate of the data and to allow for and discard the last figure so obtained.

In the cases of Tables and Graphs, of course, the limits of accuracy are clearly defined.

Errors of Functions.

Suppose A, B, $C \ldots$ are exact quantities, and that their calculated values are A', B', $C' \ldots$, where $A' = A + a$, $B' = B + b$, $C' = C + c$ and so on.

Then a, b, $c \ldots$ are the errors of A, B, $C \ldots$, and $\frac{a}{A}$, $\frac{b}{B}$, $\frac{c}{C}$ are the *Relative Errors* of A, B, $C \ldots$. In practice, when the errors are small in comparison with A, B, $C \ldots$, it is permissible to take $\frac{a}{A'}$, $\frac{b}{B'}$, and $\frac{c}{C'}$ for the relative errors. 5

Two kinds of errors are met with in practical work. It may be known that a, b, $c \ldots$ are the *actual* errors, or that they are the *maximum* numerical values of the errors, which may, however, lie anywhere between $-a$ and a, $-b$ and b, and $-c$ and c. In the latter case, the distribution of each error would be a straight line between $-a$ and a, $-b$ and b, and $-c$ and c. As we have seen, the Probable Error of each would be $\frac{1}{2}a$, $\frac{1}{2}b$, or $\frac{1}{2}c$.

In dealing with the errors of functions it is usual to consider only the Maximum errors or the Maximum Relative Errors.

Suppose $u = f(xyz)$, then we know that

$$du = \frac{\partial u}{dx} \cdot dx + \frac{\partial u}{dy} \cdot dy + \frac{\partial u}{dz} \cdot dz.$$

Taking du, dx, dy, and dz as a series of corresponding errors of u, x, y, and z respectively, it is clear that if the errors of y and z were zero, then $du = \frac{\partial u}{\partial x} \cdot dx$. In other words $\frac{\partial u}{\partial x} \cdot dx$ is the error in u due to an error in x only. It follows that *The Maximum Error of any function due to errors in the variables is the sum of the errors due to each separately.* The result in fact is précisely the same as if the errors occurred *consecutively* instead of *simultaneously.*

Example 1. *Find the percentage error in the area of a triangle calculated from the formula* $\Delta = \frac{1}{2}ab \sin C$, *when* $a = 100$, $b = 150$, *and* $C = 45°$. *The errors in a and b are in each case 1, and that in C is* $\frac{1}{2}°$.

Here, $\Delta = \frac{1}{2}\, ab \sin C$;

$\therefore \quad d\Delta = \frac{1}{2}\, b \sin C \,.\, da + \frac{1}{2}\, a \sin C \,.\, db + \frac{1}{2}\, ab \cos C \,.\, dC$

and $\quad \frac{d\Delta}{\Delta} = \frac{da}{a} + \frac{db}{b} + \cot C \,.\, dC.$

Since $\cot C = 1$, and $dC = 0{\cdot}0087$ (in radians),

$$\therefore \frac{d\Delta}{\Delta} = \tfrac{1}{100} + \tfrac{1}{150} + 0{\cdot}0087.$$

Multiplying throughout by 100 to turn Relative into Percentage errors, we have

$$\text{Percentage Error in } \Delta = 1 + 0{\cdot}67 + 0{\cdot}87$$
$$= 2{\cdot}54 \text{ per cent.}$$

A simple but important case of common occurrence is that when $u = Kx^p y^q z^r$.

In this case it easily follows that $\frac{du}{u} = p\frac{dx}{x} + q\frac{dy}{y} + r\frac{dz}{z}$.

If dashed letters are used to denote relative errors of the corresponding undashed letters, then $u' = px' + qy' + rz'$.

In the simplest case of all, namely when $u = Kxyz$.

Then $u' = x' + y' + z'$, a relation which may be conveniently expressed in words, as follows:

The Relative Error of a product is equal to the sum of the Relative Errors of its factors. 5

Example 2. *If $p = 4{\cdot}2x\sqrt{y}$, find what percentage error would be caused in p by a 3 per cent. error in both x and y.*

We may write down at once from

$$p' = x' + \tfrac{1}{2}y'$$

$$= \frac{3}{100} + \frac{1\frac{1}{2}}{100}$$

$$\therefore \quad p' = 4\tfrac{1}{2} \text{ per cent.}$$

It is worthy of notice that the relative error of the sum of any number of quantities must always lie between the greatest and least of the relative errors of the quantities.

Suppose $A, B, C \ldots N$ are the quantities and $a, b, c \ldots n$ are the corresponding errors.

Then a well-known Theorem states that

$$\frac{a+b+c+\ldots n}{A+B+C\ldots N}$$

lies between the greatest and least of the ratios

$$\frac{a}{A}, \frac{b}{B}, \ldots \frac{n}{N}.$$

The result follows at once.

If it is often necessary to find the Probable Error of the sum of two quantities, when the Probable Error of each is known. If the quantities are completely independent, and P_x, P_y are their probable errors, then the Probable Error of their sum or difference is $\sqrt{P_x^2 + P_y^2}$. If, on the other hand, there is a perfect correlation between the quantities, then the Probable Error of the sum is $P_x + P_y$ and of the difference $P_x \sim P_y$. In general, when the correlation coefficient is r, then the Probable Error of the sum is $\sqrt{P_x^2 + P_y^2 + 2rP_xP_y}$ and that of the difference, $\sqrt{P_x^2 + P_y^2 - 2rP_xP_y}$. (Appendix, XXIII.) 5

APPENDIX

x = error ; ρ = deviation.

Items.

I. To prove that $\int_0^{\infty} e^{-h^2x^2}\,dx = \frac{\sqrt{\pi}}{2h}$.

II. To deduce the normal equation, $y = \frac{h}{\sqrt{\pi}} e^{-h^2x^2}\,dx$.

III. The method of calculating the 2-Y table.

IV. The 2-Y table.

V. To deduce the standard formula,

$$h = \sqrt{\frac{n}{2\,\Sigma\,(x^2)}} = \frac{1}{\sqrt{2}}\,\frac{1}{\sigma}, \text{ or } \sqrt{\frac{(n-1)}{2\,\Sigma\,(\rho^2)}}.$$

VI. To find the mean error (M) and to deduce Peters's

$$h = \frac{1}{M\sqrt{\pi}} \text{ or } \frac{\sqrt{n\,(n-1)}}{\sqrt{\pi}\,\Sigma\,(\rho)}.$$

VII. To prove that $\sigma = \frac{1}{h\sqrt{2}}$ and that where $x = \sigma$ there is a point of inflexion on the normal curve.

VIII. To prove that the probable error, $P = \frac{0{\cdot}4769}{h} = 0{\cdot}6745\,\sigma$
$= 0{\cdot}8453\,M$.

IX. To deduce the equation of area distribution, $z = 2\,h^2 r e^{-h^2r^2}$

X. To obtain the integral equation of area distribution,

$$Z = 1 - e^{-h^2r^2}.$$

XI. To prove that in area distribution, the probable error, $Q = \frac{0{\cdot}8326}{h}$.

XII. To prove that in area distribution, $h = \frac{1}{\sigma}$, and $Q = 0{\cdot}8326\,\sigma$.

XIII. To find the mean error (M) in area distribution and to prove that $h = \frac{\sqrt{\pi}}{2\,M}$ and $Q = 0{\cdot}9394\,M$.

XIV. To deduce the equation of space distribution,

$$v = \frac{4\,h^3}{\sqrt{\pi}}\,s^2 e^{-h^2s^2} ds.$$

XV. To obtain the integral equation of space distribution,

$$V = 2Y - \frac{2h}{\sqrt{\pi}} s e^{-h^2 s^2}.$$

XVI. To prove that in space distribution, the probable error,

$$R = \frac{1 \cdot 0876}{h}.$$

XVII. To prove that in space distribution, $h = \sqrt{\frac{3}{2}} \cdot \frac{1}{\sigma}$.

XVIII. To find the mean error (M) in space distribution and to prove that $h = \frac{2}{M\sqrt{\pi}}$ and $R = 0 \cdot 9639\, M$.

XIX. Points of inflexion on the area curve.

XX. Points of inflexion on the space curve.

XXI. To prove that $\frac{\Sigma(xy)}{\sqrt{\Sigma(x^2)\,\Sigma(y^2)}}$ has a maximum value, unity, when the correlation is perfect.

XXII. To prove that $r = r'(1 + E^2)$, where r is the true correlation, the correlation uncorrected for errors of observation, and

$$E^2 = \tfrac{1}{2}\left[\left(\frac{\sigma_a}{\sigma_x}\right)^2 + \left(\frac{\sigma_b}{\sigma_y}\right)^2\right],$$

a and b being the errors of observation of x and y respectively.

XXIII. To find the probable error of the sum or difference of two quantities in terms of their separate probable errors.

XXIV. The expansions of $(1 + x + x^2 + \dots x^{10})^n$ when $n = 2, 3, 4$, or 5.

XXV. To prove that $r_{xy\cdot z} = \frac{r_{xy} - r_{xz}\cdot r_{yz}}{\sqrt{(1 - r_{xz}^2)(1 - r_{yz}^2)}}$, and a note on $r_{xy\cdot uv}$.

XXVI. To show that $\Sigma(n_c l^2) = \Sigma(y^2)(\eta_{yx}^2 - r^2)$, where l is the vertical distance from the regression line to the centre of gravity of a column of dots.

APPENDIX

The Mathematics of this Appendix has been written as simply and clearly as possible so that it may be within the range of the non-specialist. The use of double and triple integrals has been avoided.

I. *To prove that* $\int_0^\infty e^{-h^2x^2}dx = \frac{\sqrt{\pi}}{2h}$.

Let $$\int_0^\infty e^{-h^2x^2}dx = I.$$

Put $$z = hx$$

so that $$e^{-z^2}\delta z = he^{-h^2x^2}\delta x \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad \text{A}$$

Multiply both sides by $e^{-h^2}\delta h$.

Then $$e^{-z^2}\delta z \,.\, e^{-h^2}\delta h = he^{-h^2x^2}\delta x \,.\, e^{-h^2}\delta h.$$

Suppose that z, x, δz, and δx have fixed values, we may integrate both sides with respect to h.

$$\int_0^\infty e^{-z^2}\delta z \,.\, e^{-h^2}dh = \int_0^\infty \delta x \,.\, he^{-h^2(1+x^2)}dh.$$

In the integral on the R.H. side put,

$$h^2(1+x^2) = z'$$

so that $$2h(1+x^2)\delta h = \delta z'.$$

Then $$\int_0^\infty \delta x \,.\, he^{-h^2(1+x^2)}dh = \frac{\delta x}{2(1+x^2)}\int_0^\infty e^{-z'}dz'$$

$$= \frac{\delta x}{2(1+x^2)},$$

since $\int_0^\infty e^{-z'}dz' = 1$;

$$\therefore \int_0^\infty e^{-z^2}\delta z \,.\, e^{-h^2}dh = \frac{\delta x}{2(1+x^2)},$$

$$\therefore e^{-z^2}dz \,.\, hI = \frac{\delta x}{2(1+x^2)}, \quad \text{(from A)}$$

3462 F

$$\therefore \quad hI\int_0^\infty e^{-z^2}dz = \int_0^\infty \frac{dx}{2(1+x^2)},$$

$$\therefore \quad h^2I^2 = \tfrac{1}{2}\Big[\tan^{-1}x\Big]_0^\infty, \text{ since } \int_0^\infty e^{-z^2}dz = hI$$

$$= \frac{\pi}{4},$$

$$\therefore \quad hI = \frac{\sqrt{\pi}}{2},$$

$$\therefore \quad I = \frac{\sqrt{\pi}}{2h}.$$

II. *To deduce the equation,* $y = \frac{h}{\sqrt{\pi}}e^{-h^2x^2}dx$.

Suppose there are n elemental errors, each of magnitude ϵ, and that n is indefinitely large.

When $(n-m)$ errors are positive and m negative, the total error is $(n-2m)\epsilon$, and when $(n-m+1)$ errors are positive and $(m-1)$ negative, the total error is $(n-2m+2)\epsilon$. The interval of the observed errors must be 2ϵ.

Call the total errors of $(n-2m)\epsilon$ and $(n-2m+2)\epsilon$, $x-\epsilon$ and $x+\epsilon$ respectively, so that $x = (n-2m+1)\epsilon$. . . . . A

Now the number of ways in which $(n-m)$ positive errors and m negative errors can be chosen out of n errors, each of which may be positive or negative, is $\frac{n!}{(n-m)!\,m!}$.

If $f(x-\epsilon)$ is the frequency of the error $(x-\epsilon)$, then

$$f(x-\epsilon) = \frac{n!}{(n-m)!\,m!}$$

and

$$f(x+\epsilon) = \frac{n!}{(n-m+1)\,(m-1)!};$$

$$\therefore \quad \frac{f(x+\epsilon)}{f(x-\epsilon)} = \frac{m}{n-m+1}. \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad \text{B}$$

Since the errors of ϵ are assumed to be infinitesimally small in comparison with the total errors of actual measurement, the latter may be assumed to vary continuously from $-n\epsilon$ to $+n\epsilon$, and the frequency of errors lying between $(x-\epsilon)$ and $(x+\epsilon)$ may be defined to be $f(x)$.

If x is represented as a distance measured along a straight line from a fixed point in it, and if δx is the distance from $(x-\epsilon)$ to $(x+\epsilon)$, then the frequency *per unit length* is $\dfrac{f(x)}{\delta x}$.

Let $$\phi(x) = \frac{f(x)}{\delta x}$$

so that $$f(x) = \phi(x)\delta x.$$

As x is assumed to vary continuously, we may write

$$f(x) = \phi(x)dx$$

where dx is indefinitely small.

It follows that $$\frac{\phi(x+\epsilon)}{\phi(x-\epsilon)} = \frac{f(x+\epsilon)}{f(x-\epsilon)} = \frac{m}{n-m+1} \quad . \quad . \quad \text{(from B)}$$

$$\therefore \quad \frac{\phi(x+\epsilon)-\phi(x-\epsilon)}{\phi(x+\epsilon)+\phi(x-\epsilon)} = \frac{-n+2m-1}{n+1} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad \text{C}$$

Neglecting squares and higher powers of ϵ, we have:

$$\phi(x+\epsilon) = \phi(x) + \epsilon\frac{d[\phi(x)]}{dx}$$

and $$\phi(x-\epsilon) = \phi(x) - \epsilon\frac{d[\phi(x)]}{dx}.$$

Substituting these values in C, we have:

$$\frac{2\epsilon\dfrac{d[\phi(x)]}{dx}}{2\phi(x)} = -\frac{n-2m+1}{n+1} = -\frac{x}{(n+1)\epsilon}; \quad \text{(from A)}$$

$$\therefore \quad \frac{1}{\phi(x)} \cdot \frac{d[\phi(x)]}{dx} = -\frac{x}{(n+1)\epsilon^2}.$$

Integrating this equation, we have:

$$\log[\phi(x)] = -\frac{x^2}{2(n+1)\epsilon^2} + \text{constant}.$$

$$\therefore \quad \phi(x) = e^{-h^2x^2 + \text{constant}}$$

where $$h^2 = \frac{1}{2(n+1)\epsilon^2};$$

$$\therefore \quad \phi(x) = ke^{-h^2x^2}.$$

Now $\phi(x)$ is the frequency per unit length.

Hence $\phi(x)\,\delta x$ is the number of errors lying between the limits $x-\frac{1}{2}\delta x$ and $x+\frac{1}{2}\delta x$ or practically between the limits x and $x+\delta x$.

And $\frac{\phi(x)\,\delta x}{N}$ is the probability of an error lying between x and $x+\delta x$, where N is the number of total errors.

But the sum of all the probabilities is 1 ; . . . . . . (9)

$$\therefore \int_{-\infty}^{+\infty} \frac{\phi(x)\,dx}{N} = \int_{-\infty}^{+\infty} \frac{k}{N} e^{-h^2x^2} dx = 1,$$

$$\therefore 2\frac{k}{N} \cdot \frac{\sqrt{\pi}}{2h} = 1 \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad \text{(I)}$$

and

$$\frac{k}{N} = \frac{h}{\sqrt{\pi}}.$$

Writing y for $\frac{\phi(x)\,\delta x}{N}$ we have :

$$y = \frac{k}{N} \cdot e^{-h^2x^2} dx$$

$$= \frac{h}{\sqrt{\pi}} e^{-h^2x^2} dx.$$

III. *The method of calculating the 2-Y table.*

Let the total probability that an error is $< X$, some finite value of x, be Y, where

$$Y = \frac{h}{\sqrt{\pi}} \int^{X} e^{-h^2x^2} dx.$$

Put $hx = z$, so that $dx = \frac{dz}{h}$.

$$\text{Then } Y = \frac{1}{\sqrt{\pi}} \int_0^{hX} e^{-z^2} dz = \frac{1}{\sqrt{\pi}} \int_0^{hX} e^{-x^2} dx. \quad . \quad . \quad . \quad . \quad . \quad \text{A}$$

Now $\int e^{-x^2} dx$ cannot be found directly, so e^{-x^2} is split up into a series

$$e^{-x^2} = 1 - x^2 + \frac{x^4}{2\,!} - \frac{x^6}{3\,!} \cdots$$

whence,

$$\int_0^{hX} e^{-x^2}dx = hX - \frac{(hX)^3}{3} + \frac{1}{2!}\frac{(hX)^5}{5} - \frac{1}{3!}\frac{(hX)^7}{7} + \dots .$$

This converges, and is used when hX is small.

When hX is large, a different method is necessary:

$$\int e^{-x^2}dx = \int \left(-\frac{1}{2x}\right)\left(-2xe^{-x^2}dx\right) = \int \left(-\frac{1}{2x}\right)d\left(e^{-x^2}\right).$$

Integrating by parts successively, we have:

$$\int e^{-x^2}dx = -\frac{1}{2x}e^{-x^2} - \tfrac{1}{2}\int \frac{e^{-x^2}}{x^2}dx$$

$$= -\frac{1}{2x}e^{-x^2} + \frac{e^{-x^2}}{4x^3} + \tfrac{3}{4}\int \frac{e^{-x^2}}{x^4}dx,$$

since $\displaystyle\int \frac{e^{-x^2}}{x^2}dx = \int \left(-\frac{1}{2x^3}\right)\left(-2xe^{-x^2}dx\right)$

$$= \int \left(-\frac{1}{2x^3}\right)d\left(e^{-x^2}dx\right).$$

Continuing, we have:

$$\int e^{-x^2}dx$$

$$= e^{-x^2}\left[-\frac{1}{2x} + \frac{1}{4x^3} - \frac{3}{8x^5} + \frac{3.5}{16x^7} - \frac{3.5.7}{32x^9} + \frac{3.5.7.9}{64x^{11}}\dots\right]. \quad \text{B}$$

Now

$$Y = \frac{1}{\sqrt{\pi}}\int_0^{hX} e^{-x^2}dx = \frac{1}{\sqrt{\pi}}\int_0^{\infty} e^{-x^2}dx - \frac{1}{\sqrt{\pi}}\int_{hX}^{\infty} e^{-x^2}dx \quad . \text{ (A)}$$

$$= \tfrac{1}{2} - \frac{1}{\sqrt{\pi}}\int_{hX}^{\infty} e^{-x^2}dx. \left(\text{Since} \int_0^{\infty} e^{-x^2}dx = \frac{\sqrt{\pi}}{2}.\right) \quad . \quad . \text{ (I)}$$

The integral on the R.H. side of the equation can be evaluated, from series B, when $hX > 1$.*

The value of $2Y$ is more useful in practice than Y, because the normal curve is symmetrical about the y-axis, and as errors

* Series B is an *asymptotic expansion*. The numerical value of the function differs from the sum of a finite number of terms by an amount less than the last term included. Accordingly, to get the best approximation, retain the terms until they cease to diminish.

are usually considered irrespective of sign, it is more useful to take limits from $-X$ to $+X$ than it is from 0 to $+X$. $2Y$ is the fraction which the errors occurring between $-hx$ and $+hx$ form of the whole number of errors.

IV. *The 2·Y table.*

Where
$$Y = \frac{1}{\sqrt{\pi}} \int_0^{hX} e^{-x^2} dx$$

or
$$\frac{h}{\sqrt{\pi}} \int_0^{X} e^{-h^2 x^2} dx.$$

2 *Y*.

hX	**0**	1	2	3	4	**5**	6	7	8	9
0·0	0·0000	0·0113	0·0226	0·0338	0·0451	0·0564	0·0676	0·0789	0·0901	0·1013
0·1	0·1125	0·1236	0·1348	0·1459	0·1569	0·1680	0·1790	0·1900	0·2009	0·2118
0·2	0·2227	0·2335	0·2443	0·2550	0·2657	0·2763	0·2869	0·2974	0·3079	0·3183
0·3	0·3286	0·3389	0·3491	0·3593	0·3694	0·3794	0·3893	0·3992	0·4090	0·4187
0·4	0·4284	0·4380	0·4475	0·4569	0·4662	0·4755	0·4847	0·4937	0·5027	0·5117
0·5	0·5205	0·5292	0·5379	0·5465	0·5549	0·5633	0·5716	0·5798	0·5879	0·5959
0·6	0·6039	0·6117	0·6194	0·6270	0·6346	0·6420	0·6494	0·6566	0·6638	0·6708
0·7	0·6778	0·6847	0·6914	0·6981	0·7047	0·7112	0·7175	0·7238	0·7300	0·7361
0·8	0·7421	0·7480	0·7538	0·7595	0·7651	0·7707	0·7761	0·7814	0·7867	0·7918
0·9	0·7969	0·8019	0·8068	0·8116	0·8163	0·8209	0·8254	0·8299	0·8342	0·8385
1·0	0·8427	0·8468	0·8508	0·8548	0·8586	0·8624	0·8661	0·8698	0·8733	0·8768
1·1	0·8802	0·8835	0·8868	0·8900	0·8931	0·8961	0·8991	0·9020	0·9048	0·9076
1·2	0·9103	0·9130	0·9155	0·9181	0·9205	0·9229	0·9252	0·9275	0·9297	0·9319
1·3	0·9340	0·9361	0·9381	0·9400	0·9419	0·9438	0·9456	0·9473	0·9490	0·9507
1·4	0·9523	0·9539	0·9554	0·9569	0·9583	0·9597	0·9611	0·9624	0·9637	0·9649
1·5	0·9661	0·9673	0·9684	0·9695	0·9706	0·9716	0·9726	0·9736	0·9745	0·9755
1·6	0·9763	0·9772	0·9780	0·9788	0·9796	0·9804	0·9811	0·9818	0·9825	0·9832
1·7	0·9838	0·9844	0·9850	0·9856	0·9861	0·9867	0·9872	0·9877	0·9882	0·9886
1·8	0·9891	0·9895	0·9899	0·9903	0·9907	0·9911	0·9915	0·9918	0·9922	0·9925
1·9	0·9928	0·9931	0·9934	0·9937	0·9939	0·9942	0·9944	0·9947	0·9949	0·9951
2·0	0·9953	0·9955	0·9957	0·9959	0·9961	0·9963	0·9964	0·9966	0·9967	0·9969
2·1	0·9970	0·9972	0·9973	0·9974	0·9975	0·9976	0·9977	0·9979	0·9980	0·9980
2·2	0·9981	0·9982	0·9983	0·9984	0·9985	0·9985	0·9986	0·9987	0·9987	0·9988
2·3	0·9989	0·9989	0·9990	0·9990	0·9991	0·9991	0·9992	0·9992	0·9992	0·9993
2·4	0·9993	0·9993	0·9994	0·9994	0·9994	0·9995	0·9995	0·9995	0·9995	0·9996
2·5	0·999593									
2·6	0·999764									
2·7	0·999866									
2·8	0·999925									
2·9	0·999959									
	3·00 0·999978								**4·00** 0·9999999846	

V. ***To deduce the Standard Formula,***

$$h = \sqrt{\frac{n}{2\Sigma(x^2)}} = \frac{1}{\sqrt{2}} \cdot \frac{1}{\sigma} \text{ or } \sqrt{\frac{(n-1)}{2\Sigma(\rho^2)}}.$$

Let $y_1, y_2, y_3 \dots y_n$ be the probabilities of the errors

$$x_1, x_2, x_3 \dots x_n.$$

Since $y = \frac{h}{\sqrt{\pi}} e^{-h^2x^2} \delta x$, the total probability (Y) of all the errors is given by the equation $Y = \left(\frac{\delta x}{\sqrt{\pi}}\right)^n \times h^n e^{-h^2\Sigma(x^2)}$.

Regarding h as a variable and the x's and δx as constants, the most probable value of h is that for which Y is at a maximum, and $\frac{dY}{dh} = 0$.

Now $\frac{dY}{dh} = \left(\frac{\delta x}{\sqrt{\pi}}\right)^n e^{-h^2\Sigma(x^2)} \left(nh^{n-1} - 2h^{n+1}\Sigma(x^2)\right)$;

$\therefore \quad \frac{dY}{dh} = 0$, when $nh^{n-1} = 2h^{n+1}\Sigma(x^2)$,

that is, when $h = \sqrt{\frac{n}{2\Sigma(x^2)}}$.

And since $\sigma = \sqrt{\frac{\Sigma(x^2)}{n}}$,

$$h = \frac{1}{\sqrt{2}} \cdot \frac{1}{\sigma}.$$

If allowance is made for errors being replaced by deviations,

$$h = \sqrt{\frac{(n-1)}{2\Sigma(\rho^2)}}. \quad . \quad . \quad \text{(See note, p. 17.)}$$

VI. ***To find the Mean error (M) and to deduce Peters's Formula,***

$$h = \frac{1}{M\sqrt{\pi}} \text{ or } h = \frac{\sqrt{n(n-1)}}{\sqrt{\pi}\,\Sigma(\rho)}.$$

The Mean error refers to the Mean of the errors on *one* side of the curve, for the Mean of *all* the errors is assumed to be 0.

If $f_1\ f_2\ f_3$ &c. are the frequencies (that is, the actual numbers of occurrences) of $x_1\ x_2\ x_3$ &c. all of the same sign, and if N is the total number of errors:

$$\text{the Mean,}\quad M = \frac{f_1x_1 + f_2x_2 + f_3x_3 + \dots}{f_1 + f_2 + \dots}$$

$$= \frac{f_1x_1 + f_2x_2 + f_3x_3 + \dots}{\frac{N}{2}}$$

$$= 2[x_1y_1 + x_2y_2 + x_3y_3 + \dots] \quad . \quad . \quad . \quad . \quad . \quad . \quad (6)$$

$$= \frac{2h}{\sqrt{\pi}}\int_0^\infty xe^{-h^2x^2}dx.$$

Perhaps the simplest way to regard the Mean error is to look upon it as the *Centre of Gravity* of masses proportional to the probabilities (or frequencies) at corresponding x-distances from the y-axis. The Mean is the distance of the Centre of Gravity of the area of one-half of the curve from the y-axis.

$$\text{Now}\qquad \int_0^\infty xe^{-h^2x^2}dx = -\frac{1}{2h^2}\Big[e^{-h^2x^2}\Big]_0^\infty\ *$$

$$= -\frac{1}{2h^2}(0-1)$$

$$= \frac{1}{2h^2};$$

$$\therefore\quad M = \frac{1}{h\sqrt{\pi}},$$

$$\text{and}\qquad h = \frac{1}{M\sqrt{\pi}}.$$

If allowance is made for replacing errors by deviations,

$$h = \frac{\sqrt{n(n-1)}}{\sqrt{\pi}\,\Sigma(\rho)} \quad . \quad . \quad . \quad \text{(See note, p. 17.)}$$

* Easily checked by differentiating. The substitution is $z = h^2x^2$. See XI.

VII. *To prove that* $\sigma = \frac{1}{h\sqrt{2}}$ *and that where* $x = \sigma$ *there is a point of inflexion on the Normal Curve.*

Since $h = \frac{1}{\sqrt{2}} \cdot \frac{1}{\sigma}$, ... VI

$$\therefore \quad \sigma = \frac{1}{h\sqrt{2}}.$$

Since $y = ce^{-h^2x^2}$ $\left(\text{where } c = \frac{h\delta x}{\sqrt{\pi}}\right)$,

$$\therefore \quad \frac{dy}{dx} = -2ch^2xe^{-h^2x^2}.$$

$$\text{And} \quad \frac{d^2y}{dx^2} = -2ch^2\left[e^{-h^2x^2} - 2h^2x^2e^{-h^2x^2}\right]$$

$$= -2ch^2e^{-h^2x^2}(1-2h^2x^2).$$

When $$\frac{d^2y}{dx^2} = 0$$

$$1 - 2h^2x^2 = 0,$$

and $$x = \pm\frac{1}{h\sqrt{2}} = \pm\sigma.$$

Moreover, when $$x = \pm\frac{1}{h\sqrt{2}},$$

$$y = ce^{-\frac{1}{2}} = \frac{c}{\sqrt{e}}.$$

Hence, where $x = \sigma$ there is a point of **inflexion** on the Normal Curve, and at that point $y = \frac{c}{\sqrt{e}}$.

VIII. *To prove that the Probable Error,* $P = \frac{0{\cdot}4769}{h} = 0{\cdot}6745\,\sigma$
$= 0{\cdot}8453\,M.$

For definition of Probable Error see p. 20.

The value of P is that value of x for which $2\,Y = \frac{1}{2}$.

By interpolation in the $2{\cdot}Y$ table,

$$hP = 0{\cdot}4769,$$

$$\therefore \quad P = \frac{0{\cdot}4769}{h}.$$

Taking $h = \frac{1}{\sqrt{2}} \cdot \frac{1}{\sigma}$,

$$P = 0{\cdot}4769\sqrt{2}\sigma$$
$$= 0{\cdot}6745\sigma.$$

Taking $h = \frac{1}{M\sqrt{\pi}}$,

$$P = 0{\cdot}4769\sqrt{\pi}M$$
$$= 0{\cdot}8453M.$$

IX. *To deduce the equation of area-distribution,*

$$z = 2h^2 re^{-h^2r^2} dr.$$

Let O be the Mean position of the 'shots (fig. 42).

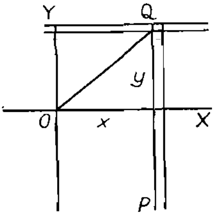


FIG. 42.

Through O draw any X-Y frame.

Take a strip PQ parallel to OY and at a distance x, and having its breadth δx.

It will be assumed that the number of shots falling on the strip is proportional to the number of shots which would fall on its width on the x-axis if the distribution were linear along the x-axis, and that the distribution along the strip itself is normal. As will be seen from what follows, these assumptions are equivalent to assuming that the distribution per unit area varies with the distance from the centre according to the normal frequency law.

Let N be the total number of shots.

Then the number falling on the strip $PQ \times \delta x$

$$= \frac{h\delta x}{\sqrt{\pi}} e^{-h^2x^2} \times N.$$

But since the distribution along PQ is normal,

∴ the number falling on $\delta x \,.\, \delta y$ at Q

$$= \frac{h\delta y}{\sqrt{\pi}} e^{-h^2y^2} \times \frac{h\delta x}{\sqrt{\pi}} e^{-h^2x^2} \times N$$

$$= \frac{h^2}{\pi} e^{-h^2(x^2+y^2)} \delta x \,.\, \delta y \times N.$$

$\therefore$ The number of shots *per unit area* at Q

$$= \frac{h^2}{\pi} e^{-h^2 r^2} \times N, \text{ where } r^2 = x^2 + y^2.$$

$\therefore$ The number of shots in a ring of radius r and width δr

$$= \frac{h^2}{\pi} \cdot 2\pi r e^{-h^2 r^2} \delta r \times N.$$

If z is the probability of shots falling in the ring,

$$\text{then } z = 2h^2 r e^{-h^2 r^2} dr.$$

X. *To obtain the integral equation of area-distribution,*

$$Z = 1 - e^{-h^2 r^2}.$$

If Z is the total probability of errors not exceeding r,

then $$Z = \int 2h^2 r e^{-h^2 r^2} dr.$$

Put $$h^2 r^2 = z,$$

so that $$2h^2 r dr = dz.$$

Then $$Z = \int e^{-z} dz$$

$$= -e^{-z} + C = -e^{-h^2 r^2} + C.$$

And when $$r = \infty, \; Z = 1,$$

$$\therefore \quad 1 = 0 + C,$$

$$\therefore \quad Z = 1 - e^{-h^2 r^2}.$$

XI. *To prove that in area-distribution the Probable Error,*

$$Q = \frac{0 \cdot 8326}{h}.$$

If Q is the probable error, then the total probability of *errors not exceeding* $Q = \frac{1}{2}$.

$$\therefore \quad \left[1-e^{-h^2r^2}\right]_0^Q = \tfrac{1}{2};$$

$$\therefore \quad 1-e^{-h^2Q^2} = \tfrac{1}{2};$$

$$\therefore \quad e^{-h^2Q^2} = \tfrac{1}{2};$$

$$\therefore \quad h^2Q^2 = \log_e 2 = 0{\cdot}69315;$$

$$\therefore \quad hQ = 0{\cdot}8326;$$

$$\therefore \quad Q = \frac{0{\cdot}8326}{h}.$$

XII. *To prove that in area-distribution* $h = \frac{1}{\sigma}$ *and* $Q = 0{\cdot}8326\sigma$.

The probabilities of the errors r_1, r_2, r_3, &c., are z_1, z_2, z_3, &c., where $z = 2h^2re^{-h^2r^2}dr$. The total probability of all the errors is $z_1z_2z_3 \ldots z_n$, which may be written $\Pi(z)$.

And $\Pi(z) = 2^n \,.\, h^{2n}r_1r_2r_3 \ldots r_n\, e^{-h^2(r_1^2+r_2^2+\ldots r_n^2)}(\delta r)^n$.

Suppose that the r's and (δr) are fixed quantities and the h is an independent variable. Differentiating with regard to h, and writing $\Pi(r)$ for $r_1r_2 \ldots r_n$ and $\Sigma(r^2)$ for $r_1^2+r_2^2+ \ldots r_n^2$ we have:

$$d[\Pi(z)]/dh$$

$$= 2^n\,\Pi(r)\,(\delta r)^n\left[2nh^{2n-1}e^{-h^2\Sigma(r^2)} - 2h^{2n+1}\,\Sigma(r^2)e^{-h^2\Sigma(r^2)}\right]$$

$$= 2^n\,\Pi(r)\,(\delta r)^n \,.\, e^{-h^2\Sigma(r^2)}\left[2nh^{2n-1} - 2h^{2n+1}\,\Sigma(r^2)\right].$$

When $\Pi(z)$ is at a maximum, the corresponding value of h is the most probable one. This will be the case when $d[\Pi(z)]/dh = 0$; that is, when $(2n \,.\, h^{2n-1} - 2h^{2n+1}\,\Sigma(r^2)) = 0$.

This will be the case when, $n - h^2\Sigma(r^2) = 0$,

that is, when $$h = \sqrt{\frac{n}{\Sigma(r^2)}},$$

or when $$h = \frac{1}{\sigma}.$$

Since $$Q = \frac{0{\cdot}8326}{h},$$

$$\therefore \quad Q = 8326\,\sigma.$$

XIII. *To find the Mean error in (M) **area-distribution and to** prove that*

$$h = \frac{\sqrt{\pi}}{2M} \quad \textit{and} \quad Q = 0{\cdot}9394\,M.$$

The Mean error is the distance of the Centre of **Gravity** of the area of the distribution curve from the y-axis. (VI)

$$\therefore \quad M = 2h^2 \int_0^\infty r^2 e^{-h^2 r^2}\,dr.$$

Now $$\int r^2 e^{-h^2r^2}\,dr = \int r\,.\,re^{-h^2r^2}\,dr = \int r\,.\,d\left(\frac{e^{-h^2r^2}}{-2h^2}\right)$$

$$= r\left(-\frac{1}{2h^2}e^{-h^2r^2}\right) + \frac{1}{2h^2}\int e^{-h^2r^2}\,dr,$$

(Integrating by parts.)

$$\therefore \quad M = \Big[-re^{-h^2r^2}\Big]_0^\infty + \int_0^\infty e^{-h^2r^2}\,dr.$$

Now $$\Big[-re^{-h^2r^2}\Big]_0^\infty = 0,$$

and $$\int_0^\infty e^{-h^2r^2}\,dr = \frac{\sqrt{\pi}}{2h} \quad \ldots\ldots\ldots \quad \text{(I)}$$

$$\therefore \quad M = \frac{\sqrt{\pi}}{2h},$$

and $$h = \frac{\sqrt{\pi}}{2M}.$$

But $$Q = \frac{0{\cdot}8326}{h} = \frac{0{\cdot}8326 \times 2}{\sqrt{\pi}}\,.\,M$$

$$= 0{\cdot}9394\,M.$$

XIV. *To deduce the equation of space-distribution,*

$$v = \frac{4h^3}{\sqrt{\pi}} s^2 e^{-h^2 s^2} ds.$$

Let O be the centre of the distribution (fig. 43).

Through O draw any frame of reference $OXYZ$.

It is assumed that the number of shots in a slice $ABCD$ parallel to the y-z plane and of thickness δx is proportional to the number of shots that would fall on the breadth δx along the x-axis, in linear distribution; and that the number of shots in a prism $EFGH$ of depth δz in the slice $ABCD$ is proportional to the number of shots that would fall on δz if all the shots in the slice were on a vertical line parallel to OZ, the central position being the point where OX strikes the plane $ABCD$. Moreover, in the prism itself it is assumed that the shots are distributed normally, the central position being the point where the aforesaid vertical line strikes the prism. It is evident from what follows that these assumptions are equivalent to the assumption that the density per unit volume varies with the distance from the centre according to the normal law.

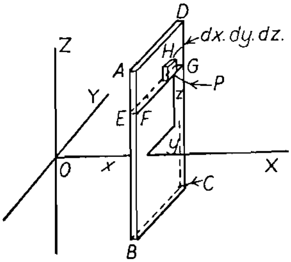


FIG. 43.

Let N be the total number of shots:

Then the number falling in the slice $ABCD \times \delta x$

$$= \frac{h\delta x}{\sqrt{\pi}} e^{-h^2 x^2} \times N,$$

and the number falling in the prism $EFGH \times \delta z$

$$= \frac{h\delta x}{\sqrt{\pi}} e^{-h^2 x^2} \times \frac{h\delta z}{\sqrt{\pi}} e^{-h^2 z^2} \times N,$$

and the number falling in the little box-shape, $\delta x \,.\, \delta y \,.\, \delta z$ at $P\ (x, y, z)$

$$=\frac{h\delta x}{\sqrt{\pi}}e^{-h^2x^2}\times\frac{h\delta z}{\sqrt{\pi}}e^{-h^2z^2}\times\frac{h\delta y}{\sqrt{\pi}}e^{-h^2y^2}\times N;$$

the number of shots at P *per unit volume*

$$=\frac{h^3}{\pi^{\frac{3}{2}}}\cdot e^{-h^2s^2}\times N \quad (\text{where } x^2+y^2+z^2=s^2);$$

the number of shots in a shell of radius s and width δs

$$=\frac{h^3}{\pi^{\frac{3}{2}}}\cdot 4\pi s^2\cdot e^{-h^2s^2}\delta s\times N.$$

Hence if v is the probability of a shot falling in this shell,

$$v=\frac{4h^3}{\sqrt{\pi}}s^2e^{-h^2s^2}ds.$$

XV. *To obtain the integral equation of space-distribution,*

$$V=2Y-\frac{2h}{\sqrt{\pi}}se^{-h^2s^2}.$$

Let V be the probability of all the errors which do not exceed v:

Then $V=\int\frac{4h^3}{\sqrt{\pi}}s^2e^{-h^2s^2}ds=\int\frac{4h^3}{\sqrt{\pi}}\cdot s\cdot se^{-h^2s^2}ds$

$$=-\frac{2h}{\sqrt{\pi}}\int s\cdot d(e^{-h^2s^2}).$$

Integrating by parts, we have:

$$V=-\frac{2h}{\sqrt{\pi}}\left[se^{-h^2s^2}-\int e^{-h^2s^2}ds\right],$$

$$=-\frac{2h}{\sqrt{\pi}}se^{-h^2s^2}+\frac{2h}{\sqrt{\pi}}\int e^{-h^2s^2}ds,$$

$$=-\frac{2h}{\sqrt{\pi}}se^{-h^2s^2}+2Y+C. \qquad \text{(III, IV)}$$

Proceeding from 0 to ∞, $\left[V\right]_0^\infty=1$, ... (9)

$\left[2Y\right]_0^\infty=1$, and $\left[se^{-h^2s^2}\right]_0^\infty=0$; $\therefore\ C=0$;

$$V=+2Y-\frac{2h}{\sqrt{\pi}}se^{-h^2s^2}.$$

XVI. *To prove that in space-distribution the Probable Error,*

$$R = \frac{1{\cdot}0876}{h}.^{*}$$

R is the value of s for which $V = \frac{1}{2}$.

It is necessary to interpolate in the numerical values of V.

Tabulate $\frac{1}{2}V$ and having found the values of s between which R lies, interpolate. $\frac{1}{2}V$ should be 0·25.

hs	h^2s^2	$\frac{hs}{\sqrt{\pi}}$	$e^{h^2s^2}$	$e^{-h^2s^2}$	$\frac{hs}{\sqrt{\pi}}e^{-h^2s^2}$	Y_s	$\frac{1}{2}V$
1·08	1·1664	0·6093	3·210	0·3115	0·1897	0·4366	0·2469
1·09	1·1881	0·6149	3·281	0·3048	0·1874	0·4384	0·2510
1·10	1·21	0·6206	3·353	0·2982	0·1850	0·4401	0·2551

Whence by interpolation $hR = 1{\cdot}0876$;

$$\therefore\ R = \frac{1{\cdot}0876}{h}.$$

XVII. *To prove that in space-distribution* $h = \sqrt{\frac{3}{2}}\cdot\frac{1}{\sigma}$,

and $R = 0{\cdot}8880\,\sigma$.*

Since $$v = \frac{4h^3}{\sqrt{\pi}}s^2e^{-h^2s^2}ds.$$

The total probability of all the errors occurring is $\Pi(v)$ . . (XII)

where $\Pi(v) \equiv v_1v_2v_3 \dots v_n$.

Regarding the s's and (δs) as fixed quantities we may differentiate with respect to h.

Then $d[\Pi(v)]/dh$

$$= \frac{4^n}{\pi^{\frac{n}{2}}}\Pi(s^2)(\delta s)^n\left[3nh^{3n-1}e^{-h^2\Sigma(s^2)} - 2h^{3n+1}\Sigma(s^2)e^{-h^2\Sigma(s^2)}\right].$$

When $\Pi(v)$ is a maximum, the corresponding value of h is the most probable one.

This will be the case when $d[\Pi(v)]/dh = 0$, that is, when $(3n\,.\,h^{3n-1} - 2h^{3n+1}\,.\,\Sigma(s^2)) = 0$.

* The last digit is approximate.

That is, when $3n = 2h^2\,\Sigma(s^2)$,

and $$h = \sqrt{\frac{3}{2}} \cdot \frac{1}{\sigma}.$$

Since, $$R = \frac{1{\cdot}0876}{h};$$

$$\therefore \quad R = 1{\cdot}0876\sqrt{\frac{2}{3}} \cdot \sigma;$$

$$= 0{\cdot}8880\,\sigma.^*$$

XVIII. *To find the Mean error (M) in space-distribution and to show that*

$$h = \frac{2}{M\sqrt{\pi}}, \quad \text{and} \quad R = 0{\cdot}9639\,M.^*$$

$$M = \int_0^\infty \frac{4h^3}{\sqrt{\pi}} s^3 e^{-h^2s^2} ds \quad . \quad . \quad . \quad . \quad . \quad \text{(VI, XIII)}$$

$$= -\frac{2h}{\sqrt{\pi}}\int_0^\infty s^2 . \, d\left(e^{-h^2s^2}\right);$$

$$= -\frac{2h}{\sqrt{\pi}}\left[s^2 . \, e^{-h^2s^2} - \int 2se^{-h^2s^2}ds\right]_0^\infty \quad . \quad . \quad . \quad .$$

(integrating by parts);

$$= -\frac{2}{h\sqrt{\pi}}\left[h^2s^2e^{-h^2s^2} + e^{-h^2s^2}\right]_0^\infty$$

$$= -\frac{2}{h\sqrt{\pi}}\left[(0+0)-(0+1)\right];$$

$$\therefore \; M = \frac{2}{h\sqrt{\pi}},$$

and $$h = \frac{2}{M\sqrt{\pi}},$$

but $$R = \frac{1{\cdot}0876}{h} = \frac{1{\cdot}0876\sqrt{\pi}}{2}M$$

$$= 0{\cdot}9639\,M.$$

* The last digit is approximate.

XIX. *Points of Inflexion on the Area-curve.*

For any real distribution h^2 is constant, and δr is the common width of the rings. Hence, it is permissible to write k for $2h^2 . \delta r$.

Differentiate the equation $z = kre^{-h^2r^2}$ with respect to r. This operation is equivalent to finding the slope of the tangent on the curve already drawn from the finite 'δr', the width of the rings. The 'dr' now dealt with has nothing to do with this constant 'δr', which has been used and is finished with.

$$\frac{dz}{dr} = k\left[e^{-h^2r^2} - 2h^2r^2e^{-h^2r^2}\right] = ke^{-h^2r^2}(1-2h^2r^2).$$

There will be a *maximum* where $\frac{dz}{dr} = 0$, that is, where

$$(1-2h^2r^2) = 0 \quad \text{and} \quad r = \frac{1}{h\sqrt{2}}. \quad . \quad . \quad . \quad . \quad \text{A}$$

Again $\frac{d^2z}{dr^2} = k(-2h^2re^{-h^2r^2})(1-2h^2r^2) + (-4h^2r) . ke^{-h^2r^2}$;

$$= ke^{-h^2r^2}\left[-2h^2r(1-2h^2r^2) - 4h^2r\right];$$

$$= ke^{-h^2r^2}\left[-6h^2r + 4h^4r^3\right].$$

There will be a point of inflexion where $\frac{d^2z}{dr^2} = 0$, that is, where $(-6h^2r + 4h^4r^3) = 0$.

That is, where $-3 + 2h^2r^2 = 0$

and $$r = \sqrt{\frac{3}{2}} \cdot \frac{1}{h}. \quad . \quad . \quad . \quad . \quad . \quad . \quad \text{B}$$

XX. *Points of inflexion on the Space-curve.*

As in the case of the area-curve, we may write K for

$$\frac{4h^3}{\sqrt{\pi}} \cdot \delta s,$$

for h is constant for any particular distribution, and δs is the common thickness of the shells.

Remarks on the differentiation in the case of area-distribution apply equally to this case.

We have, $$v = Ks^2 e^{-h^2 s^2};$$

$$\therefore \quad \frac{dv}{ds} = 2Kse^{-h^2 s^2}(1-h^2 s^2).$$

There will be a maximum where $\frac{dv}{ds} = 0$, that is, where $(1-h^2s^2) = 0$, and $s = \frac{1}{h}$. . . . . A

Again, $$\frac{d^2v}{ds^2} = \left[2Ke^{-h^2s^2} + 2Ks(-2h^2se^{-h^2s^2})\right](1-h^2s^2)$$
$$-2h^2s \times 2Kse^{-h^2s^2}$$
$$= 2Ke^{-h^2s^2}\left[(1-2h^2s^2)(1-h^2s^2) - 2h^2s^2\right];$$
$$= 2Ke^{-h^2s^2}[1-5h^2s^2+2h^4s^4].$$

There will be points of inflexion, where $\frac{d^2v}{ds^2} = 0$, that is, where $(1-5h^2s^2+2h^4s^4) = 0$.

That is, when $$h^2s^2 = \frac{5 \pm \sqrt{17}}{4},$$

and $$hs = 1{\cdot}5102 \text{ and } 0{\cdot}4682,$$

and $$s = \frac{1{\cdot}5102}{h} \text{ and } \frac{0{\cdot}4682}{h}. \quad . \; . \; . \; . \; . \; \text{—}$$

XXI. *To prove that* $\frac{\Sigma(xy)}{\sqrt{\Sigma(x^2)\,\Sigma(y^2)}}$ *has a maximum value, **unity**, when the correlation is perfect.*

Consider the function $\Sigma(x^2)\,\Sigma(y^2) - [\Sigma(xy)]^2$ when there are three values only of each letter:

$$\Sigma(x^2)\,\Sigma(y^2) - [\Sigma(xy)]^2$$
$$= (x_1^2 + x_2^2 + x_3^2)(y_1^2 + y_2^2 + y_3^2) - (x_1y_1 + x_2y_2 + x_3y_3)^2.$$
$$= x_1^2y_1^2 + x_1^2y_2^2 + x_1^2y_3^2$$
$$+ x_2^2y_1^2 + x_2^2y_2^2 + x_2^2y_3^2$$
$$+ x_3^2y_1^2 + x_3^2y_2^2 + x_3^2y_3^2$$
$$- x_1^2y_1^2 - x_2^2y_2^2 - x_3^2y_3^2$$
$$- 2x_1x_2y_1y_2 - 2x_1x_3y_1y_3 - 2x_2x_3y_2y_3$$

$$= (x_1^2y_2^2 - 2x_1x_2y_1y_2 + x_2^2y_1^2)$$
$$+ (x_1^2y_3^2 - 2x_1x_3y_1y_3 + x_3^2y_1^2)$$
$$+ (x_2^2y_3^2 - 2x_2x_3y_2y_3 + x_3^2y_2^2)\,;$$
$$= (x_1y_2 - x_2y_1)^2 + (x_1y_3 - x_3y_1)^2 + (x_2y_3 - x_3y_2)^2.$$

If a fourth point x_4y_4 be added, we have, taking the sign Σ to refer to the first three points as above :

$$[\Sigma(x^2) + x_4^2][\Sigma(y^2) + y_4^2] - [\Sigma(xy) + x_4y_4]^2$$
$$= \Sigma(x^2)\,\Sigma(y^2) - [\Sigma(xy)]^2$$
$$+ x_4^2(y_1^2 + y_2^2 + y_3^2) + y_4^2(x_1^2 + x_2^2 + x_3^2) + x_4^2y_4^2$$
$$-2x_4y_4(x_1y_1 + x_2y_2 + x_3y_3) - x_4^2y_4^2\,;$$
$$= \Sigma(x^2)\,\Sigma(y^2) - [\Sigma(xy)]^2$$
$$+ (x_4y_1 - x_1y_4)^2 + (x_4y_2 - x_2y_4)^2 + (x_4y_3 - x_3y_4)^2$$
$$= (x_1y_2 - x_2y_1)^2 + (x_1y_3 - x_3y_1)^2 + (x_2y_3 - x_3y_2)^2$$
$$+ (x_4y_1 - x_1y_4)^2 + (x_4y_2 - x_2y_4)^2 + (x_4y_3 - x_3y_4)^2.$$

It can readily be seen that by adding one point at a time the Theorem can be shown to be true for any number of points.

Now $\Sigma(x^2)\,\Sigma(y^2) = [\Sigma(xy)]^2$, when, and only when, each bracket separately $= 0$;

that is, when $$\frac{x_1}{y_1} = \frac{x_2}{y_2} = \frac{x_3}{y_3}, \text{ \&c.}$$

And this is the condition that all the points lie on a straight line passing through the Origin ; that the correlation is perfect.

Moreover, the expression $\dfrac{\Sigma(xy)}{\sqrt{\Sigma(x^2)\,\Sigma(y^2)}}$ may be written

$$\sqrt{\frac{[\Sigma(xy)]^2}{[\Sigma(xy)]^2 + D^2}},$$

where $D^2 = \Sigma(x^2)\,\Sigma(y^2) - [\Sigma(xy)]^2$, and since $[\Sigma(xy)]^2$ and D^2 are both positive, the latter expression has a maximum value, 1, when $D = 0$.

Hence when the correlation is perfect $\dfrac{\Sigma(xy)}{\sqrt{\Sigma(x')\,\Sigma(y')}}$ is at its maximum value, 1.

XXII. *To prove that* $r = r'(1+E^2)$, *where* r *is the true correlation,* r' *is the correlation uncorrected for errors of observation, and* $E^2 = \frac{1}{2}\left[\left(\frac{\sigma_a}{\sigma_x}\right)^2 + \left(\frac{\sigma_b}{\sigma_y}\right)^2\right]$, *a being the error of observation of* x, *and* b *that of* y.

We have:

$$r = \frac{\Sigma(xy)}{\sqrt{\Sigma(x^2)\,\Sigma(y^2)}} \quad \text{and} \quad r' = \frac{\Sigma(x+a)(y+b)}{\sqrt{\Sigma(x+a)^2\,\Sigma(y+b)^2}}.$$

Now $\Sigma(x+a)(y+b) = \Sigma(xy) + \Sigma(ay) + \Sigma(bx) + \Sigma(ab)$. And since the a's and b's are entirely adventitious and independent, and promiscuous as to signs, we may assume that $\Sigma(ay)$, $\Sigma(bx)$, and $\Sigma(ab)$ tend to zero, and that $\Sigma(xy) \simeq \Sigma(x+a)(y+b)$.

Again $\Sigma(x+a)^2 = \Sigma(x^2) + 2\Sigma(ax) + \Sigma(a^2)$

$$\simeq \Sigma(x^2) + \Sigma(a^2).$$

Similarly $\Sigma(y+b)^2 \simeq \Sigma(y^2) + \Sigma(b^2)$.

Hence $$\frac{r}{r'} = \sqrt{\frac{(\Sigma(x^2)+\Sigma(a^2))(\Sigma(y^2)+\Sigma(b^2))}{\Sigma(x^2)\,\Sigma(y^2)}}$$

$$= \sqrt{\left[1+\left(\frac{\sigma_a}{\sigma_x}\right)^2\right]\left[1+\left(\frac{\sigma_b}{\sigma_y}\right)^2\right]};$$

where $\sigma_x = \sqrt{\frac{\Sigma(x^2)}{n}}$, &c.

If it is assumed that $\frac{\sigma_a}{\sigma_x}$, $\frac{\sigma_b}{\sigma_y}$ are small, the expression may be expanded as far as the first term, so that,

$$r \simeq r'\left[1+\tfrac{1}{2}\left\{\left(\frac{\sigma_a}{\sigma_x}\right)^2 + \left(\frac{\sigma_b}{\sigma_y}\right)^2\right\}\right]$$

$$\simeq r'(1+E^2), \text{ where } E^2 = \tfrac{1}{2}\left\{\left(\frac{\sigma_a}{\sigma_x}\right)^2 + \left(\frac{\sigma_b}{\sigma_y}\right)^2\right\}.$$

And if it is further assumed that $\frac{\sigma_a}{\sigma_x} = \frac{\sigma_b}{\sigma_y}$, then

$$E^2 = \left(\frac{\sigma_a}{\sigma_x}\right)^2 \text{ and } E = \frac{\sigma_a}{\sigma_x} = \frac{\sigma_b}{\sigma_y}.$$

XXIII. *To find the Probable Error of the sum or difference of two quantities from the Probable Errors of the two quantities.*

Suppose there are two measured quantities X and Y subject to errors $x_1 x_2 x_3 \ldots$ and $y_1 y_2 y_3 \ldots$ respectively.

Let P_x, P_y, and P_s be the Probable Errors in X, Y, and $(X+Y)$.

$$\text{Then}\quad P_x = \frac{0{\cdot}6745}{\sqrt{n}} \sqrt{\Sigma(x^2)} \quad \ldots\ldots\ldots \quad \text{(VIII)}$$

$$P_y = \frac{0{\cdot}6745}{\sqrt{n}} \sqrt{\Sigma(y^2)},$$

$$\text{and}\quad P_s = \frac{0{\cdot}6745}{\sqrt{n}} \sqrt{\Sigma(x+y)^2}.$$

$$\text{Now } \Sigma(x+y)^2 = \Sigma(x^2) + \Sigma(y^2) + 2\Sigma(xy),$$

$$\text{and}\quad \Sigma(xy) = r\sqrt{\Sigma(x^2)\,\Sigma(y^2)} \quad \ldots\ldots\ldots \quad (41)$$

$$\therefore\quad P_s = \frac{0{\cdot}6745}{\sqrt{n}} \sqrt{\Sigma(x^2) + \Sigma(y^2) + 2r\sqrt{\Sigma(x^2)\,\Sigma(y^2)}};$$

$$= \sqrt{(P_x)^2 + (P_y)^2 + 2r P_x . P_y}.$$

And if P_d is the Probable Error of the difference between X and Y

$$P_d = \sqrt{(P_x)^2 + (P_y)^2 - 2rP_x P_y}.$$

When the quantities X and Y are quite independent, $r = 0$,

$$\text{and}\quad P_s = P_d = \sqrt{(P_x)^2 + (P_y)^2}.$$

When the correlation is perfect, $r = 1$,

$$\text{and}\quad P_s = P_x + P_y,$$

$$\text{and}\quad P_d = P_x \sim P_y.$$

XXIV. *The expansions of* $(1 + x + x^2 + \ldots x^{10})^n$ *when* $n = 2, 3, 4,$ *and* 5 *and some deductions.*

The coefficients in the expansions are given up to and including the middle term, which is the last one in each case.

$n = 2.$

1 2 3 4 5 6 7 8 9 10 11

$n = 3.$

1 3 6 10 15 21 28 36 45 55 66 75 82 87 90 91

$n = 4.$

1 4 10 20 35 56 84 120 165 220 286 360 439 520 600
676 745 804 850 880 891

$n = 5.$

1 5 15 35 70 126 210 330 495 715 1001 1360 1795 2305
2885 3526 4215 4935 5665 6380 7051 7645 8135 8500 8725 8801

When $n = 6$, the 7 terms preceding and including the middle one are: 67662, 73578, 78777, 83062, 86262, 88242, and 88913.

The coefficients in the expansion of $(1 + x + x^2 + \ldots x^{20})^3$ are, up to the middle term :

1 3 6 10 15 21 28 36 45 55 66 78 91 105 120 136 153
171 190 210 231 250 267 282 295 306 315 322 327 330 331

The Probable Errors are found by estimating half the areas of the curves, on one side of the y-axis.

XXV. *To prove that* $r_{xy \cdot z} = \dfrac{r_{xy} - r_{xz} \cdot r_{yz}}{\sqrt{(1 - r_{xz}^2)(1 - r_{yz}^2)}}$, *and a note on* $r_{xy \cdot uv}$.

Let $x' = x - cz$ and $y' = y - dz$ where x' and y' are the values of x and y corrected, as far as possible, for the effect of z.

Then the partial correlation coefficient between x and y when the effect of z is eliminated, is the correlation coefficient between x' and y'.

Now $\Sigma (x'^2) = \Sigma (x - cz)^2 = \Sigma (x^2) - 2c \, \Sigma (xz) + c^2 \, \Sigma (z^2)$,

but $$c = \frac{\Sigma (xz)}{\Sigma (z^2)};$$

$$\therefore \; \Sigma (x'^2) = \Sigma (x^2) - 2 \frac{[\Sigma (xz)]^2}{\Sigma (z^2)} + \frac{[\Sigma (xz)]^2}{\Sigma (z^2)};$$

$$= \Sigma (x^2) - \frac{[\Sigma (xz)]^2}{\Sigma (z^2)};$$

$$= \Sigma (x^2) (1 - r_{xz}^2).$$

Similarly $\Sigma(y'^2) = \Sigma(y^2)(1 - r_{yz}^2)$;

$$\therefore \sqrt{\Sigma(x'^2)\,\Sigma(y'^2)} = \sqrt{\Sigma(x^2)\,\Sigma(y^2)}\,\sqrt{(1-r_{xz}^2)(1-r_{yz}^2)} \quad . \quad \text{A}$$

Again, $\Sigma(x'y') = \Sigma(x - cz)(y - dz)$

$$= \Sigma(xy) - c\Sigma(yz) - d\Sigma(xz) + cd\Sigma(z^2),$$

but $\quad c = \dfrac{\Sigma(xz)}{\Sigma(z^2)}$ and $d = \dfrac{\Sigma(yz)}{\Sigma(z^2)}$;

$$\therefore \Sigma(x'y') = \Sigma(xy) - \frac{\Sigma(xz)\,\Sigma(yz)}{\Sigma(z^2)} - \frac{\Sigma(xz)\,\Sigma(yz)}{\Sigma(z^2)} + \frac{\Sigma(xz)\,\Sigma(yz)}{\Sigma(z^2)};$$

$$= \Sigma(xy) - \frac{\Sigma(xz)\,\Sigma(yz)}{\Sigma(z^2)};$$

$$= \sqrt{\Sigma(x^2)\,\Sigma(y^2)}\,(r_{xy} - r_{xz}\,.\,r_{yz}) \quad . \quad . \quad . \quad . \quad . \quad \text{B}$$

$$\therefore r_{xz\cdot y} = \frac{\Sigma(x'y')}{\sqrt{\Sigma(x'^2)\,\Sigma(y'^2)}} = \frac{\sqrt{\Sigma(x^2)\,\Sigma(y^2)}\,(r_{xy} - r_{xz}\,.\,r_{yz})}{\sqrt{\Sigma(x^2)\,\Sigma(y^2)}\,\sqrt{(1-r_{xz}^2)(1-r_{yz}^2)}}$$

$$= \frac{r_{xy} - r_{xz}\,.\,r_{yz}}{\sqrt{(1-r_{xz}^2)(1-r_{yz}^2)}}.$$

Note. The proof of the formula,

$$r_{xy\cdot uv} = \frac{r_{xy\cdot v} - r_{xu\cdot v}\,.\,r_{yu\cdot v}}{\sqrt{(1-r_{xu\cdot v}^2)(1-r_{yu\cdot v}^2)}}$$

is extremely complicated. If we assume, however, that when the effect of v has been eliminated from the correlations r_{xy}, r_{xu}, and r_{yu}, these corrected correlations can then be treated as if *v had never existed*, the formula for $r_{xy\cdot uv}$ follows at once from the former, by replacing r_{xy}, r_{xu}, and r_{yu} by $r_{xy\cdot v}$, $r_{xu\cdot v}$, and $r_{yu\cdot v}$, and changing z into u. In any case this is a useful mnemonic.

XXVI. *To show that* $\Sigma(n_c l^2) = \Sigma(y^2)(\eta_{yx}^2 - r^2)$, *where l is the vertical distance from the regression line to the Centre of Gravity of a column of dots.*

Since $\quad l = m_y - ax$

$$\therefore \quad l^2 = m_y^2 + a^2x^2 - 2ax\,.\,m_y.$$

And $\Sigma(n_c l^2) = \Sigma(n_c m_y^2) + a^2 \Sigma(n_c x^2) - 2a\Sigma(x \cdot n_c m_y)$

$$= \Sigma(n_c m_y^2) + a^2 \Sigma(x^2) - 2a\Sigma(xy)$$

since $\Sigma(n_c x^2) = \Sigma(x^2)$

and $n_c m_y = \Sigma(y)$ in each column.

But $\Sigma(n_c m_y^2) = \eta_{yx}{}^2 \Sigma(y^2)$

and $a = \dfrac{\Sigma(xy)}{\Sigma(x^2)}$

$$\therefore \quad \Sigma(n_c l^2) = \eta^2_{yx}\Sigma(y^2) + \frac{[\Sigma(xy)]^2}{[\Sigma(x^2)]^2} \cdot \Sigma(x^2) - 2 \cdot \frac{\Sigma(xy)}{\Sigma(x^2)} \cdot \Sigma(xy).$$

But $\dfrac{[\Sigma(xy)]^2}{\Sigma(x^2)\,\Sigma(y^2)} = r^2$

$$\therefore \quad \Sigma(n_c l^2) = \Sigma(y^2)(\eta_{yx}{}^2 - r^2).$$

SYMBOLS

Numbers refer to pages.

f, frequency, 8

x, error, 1

ρ, deviation from the mean, usually assumed equal to x, 9

M, mean error, 16

M_x, the mean of the X's, 41, 43

M_y, the mean of the Y's, 41, 43

p, probability, 13

P, probable error, 20

P_m, probable error of the mean, 21

P_r, probable error of a correlation coefficient, 51

P_η, probable error of a correlation ratio, 69

P, probability of n independent events happening simultaneously (on *pp.* 13, 14 *only*)

$d:1$, odds in favour, 14

$d':1$, odds against, 14

y', frequency of errors between $x - \frac{1}{2}\delta x$ and $x + \frac{1}{2}\delta x$, or between x and $x + \delta x$, 9, 16

y, probability of an error between x and $x + \delta x$, 16, 84

X, Y, two measured quantities, 41

Y, the sum of the errors not exceeding x, 23

σ, standard deviation, 20

Σ, sign for 'sum of'

Π, sign for 'product of'

$\risingdotseq$, sign for approximate equality

n, number of observations, or errors

n_r, number of dots in a dot-diagram lying in a row, 53

n_c, number of dots in a dot-diagram lying in a column, 53

$-h^2$, the coefficient of x^2 in the normal equations

z', the frequency of errors lying in an annulus of radii r and $r + \delta r$, 26

z, the probability of errors lying in an annulus of radii r and $r + \delta r$, 27, 91

Z, the sum of the errors not exceeding r in area distribution, 27, 91

r, an error in area distribution, 27, 90

r', value of r uncorrected for errors of observation

$r = r_{xy} = r_{yx}$, correlation coefficient between x and y, 48, 56

$r_{xy.z}$, partial correlation coefficient of x and y when the effect of z is eliminated, 56, 103

$r_{xy.uv}$, partial correlation coefficient of x and y when the effects of u and v are eliminated, 58, 104

η_{xy}, correlation ratio of x from y, 66

η_{yx}, correlation ratio of y from x, 66

r, correlation coefficient uncorrected for errors of observation, 60, 101

Q, probable error in area distribution, 29, 91

s, error in space distribution, 29, 94

v', frequency in space distribution, 31

v, probability of an error lying in a spherical shell of radii s, and $s+\delta s$, in space distribution, 29, 94

V, the sum of the errors not exceeding v, 29, 95

$\sim$ signifies the difference between two quantities, irrespective of sign, 78, 102

δx, a finite small increment of x

dx, an infinitely small increment of x

Δ, a finite small increment of, 47

$\frac{\partial x}{\partial u}$, partial differential coefficient of u with respect to x

C, or K, an unknown constant, 63, 91, 95

$f(x)$ or $\phi(x)$, any function of x

R, probable error in space distribution

M, mean error, 16

E, observation error in correlation, 61, 101

α, β, the angles which the regression lines $y = ax$ and $x = by$ make with the axes of x and y respectively, 45, 48

a, b. See above

ϕ, total mean square contingency in Quality correlation, 63

c, coefficient of mean square contingency in Quality correlation, 63

σ_x, the Standard Deviation of x, 20

FORMULAE

Marginal reference numbers are indicated by Arabic numerals, and items in the Appendix by Roman numerals.

$f = yn$ 6.

$P = p_1 p_2 p_3 \ldots . p_n$ 7.

Maximum value of $p = 1$ 9.

$y = \frac{h}{\sqrt{\pi}} e^{-h^2 x^2} dx$ 10, II.

$h = \sqrt{\frac{n}{2 \Sigma (x^2)}} = \frac{1}{\sqrt{2}} \cdot \frac{1}{\sigma} = \frac{1}{M \sqrt{\pi}}$ (linear distribution) 12, 12 a, 13, V, VI.

$h = \frac{\sqrt{\pi}}{2M} = \frac{1}{\sigma}$ (area distribution) 23, 24, XII, XIII.

$h = \frac{2}{M\sqrt{\pi}} = \sqrt{\frac{3}{2}} \cdot \frac{1}{\sigma}$ (volume distribution) 31, 32, XVII, XVIII.

$\sigma = \sqrt{\frac{\Sigma (x^2)}{n}}$ 13.

$p = \frac{a}{a+b}$ (definition) 5.

$d = \frac{p}{1-p}, \quad d' = \frac{1-p}{p}$ 8 a.

$P = \frac{0{\cdot}4769}{h}$ 15, VIII.

$P = 0{\cdot}6745 \sqrt{\frac{\Sigma (x^2)}{n}} = 0{\cdot}6745\, \sigma$ 16, VIII.

$P = 0{\cdot}8453 \frac{\Sigma (x)}{n} = 0{\cdot}8453\, M$ 17, VIII.

$P_m = \frac{P}{\sqrt{n}}$ 19.

$P = p_1 p_2 p_3 \ldots p_n$ (probabilities) 7.

$P_{x+y} = \sqrt{(P_x)^2 + (P_y)^2 + 2 r P_x . P_y}$, &c. 56, XXIII.

$Y = \frac{h}{\sqrt{\pi}} \int_0^x e^{-h^2 x^2} dx$ 20.

$z = 2 h^2 r e^{-h^2 r^2} dr$ 21, IX.

$$Z = 2h^2\int_0^r r e^{-h^2r^2}\,dr = 1 - e^{-h^2r^2} \quad 22,\ \text{X.}$$

$$h = \frac{\sqrt{\pi}}{2M}\ \text{(area distribution)} \quad 23,\ \text{XIII.}$$

$$h = \frac{1}{\sigma}\ \text{(area distribution)} \quad 24,\ \text{XII.}$$

$$r = \sqrt{\frac{3}{2}}\cdot\frac{1}{h}\ \text{(point of inflexion on area curve)} \quad 25,\ \text{XIX.}$$

$$Q = \frac{0{\cdot}8326}{h} \quad 26,\ \text{XI.}$$

$$Q = 0{\cdot}8326\,\sigma \quad 27,\ \text{XII.}$$

$$Q = 0{\cdot}9394\,M \quad 28,\ \text{XIII.}$$

$$v = \frac{4h^3}{\sqrt{\pi}}\,s^2 e^{-h^2s^2}\,ds \quad 30,\ \text{XIV.}$$

$$R = \frac{1{\cdot}0876}{h} \quad 33,\ \text{XVI.}$$

$$R = 0{\cdot}8880\sigma \quad 34,\ \text{XVII.}$$

$$V = 2Y - \frac{2hs}{\sqrt{\pi}}\,e^{-h^2s^2} \quad 36,\ \text{XV.}$$

$$b = \frac{\Sigma(xy)}{\Sigma(y^2)},\quad a = \frac{\Sigma(xy)}{\Sigma(x^2)} \quad 39.$$

$$\tan\alpha\tan\beta = \frac{[\Sigma(xy)]^2}{\Sigma(x^2)\,\Sigma(y^2)} \quad 40.$$

$$r = r_{xy} = r_{yx} = \frac{\Sigma(xy)}{\sqrt{\Sigma(x^2)\,\Sigma(y^2)}} \quad 41.$$

$$E^2 = \frac{1-r'}{r'},\quad r = r'(1+E^2) \quad 44,\ \text{XXII.}$$

$$\eta_{yx} = \sqrt{\frac{\Sigma(n_c\,.\,m_y^{\,2})}{\Sigma(n_r\,.\,y^2)}} = \sqrt{\frac{\Sigma(n_c\,.\,m_y^{\,2})}{\Sigma(y^2)}} \quad 46.$$

$$\eta_{xy} = \sqrt{\frac{\Sigma(n_r\,.\,m_x^{\,2})}{\Sigma(n_c\,.\,x^2)}} = \sqrt{\frac{\Sigma(n_r\,.\,m_x^{\,2})}{\Sigma(x^2)}} \quad 46.$$

$$P_r = 0{\cdot}6745\,\frac{1-\rho^2}{\sqrt{n}} \quad \text{page 51.}$$

$$P_\eta = 0{\cdot}6745\,\frac{1-\eta^2}{\sqrt{n}} \quad \text{page 69.}$$

$$r_{zx\cdot y} = \frac{r_{zx} - r_{zy}\,.\,r_{xy}}{\sqrt{(1-r_{zy}^{\,2})(1-r_{xy}^{\,2})}} \quad 42,\ \text{XXV.}$$

$$r_{xy.uv} = \frac{r_{xy.v} - r_{ux.v} \times r_{uy.v}}{\sqrt{(1 - r_{ux.v}^2)(1 - r_{uy.v}^2)}} \quad 43, \textbf{XXV.}$$

$$\int_0^\infty e^{-h^2x^2}\,dx = \frac{\sqrt{\pi}}{2h} \quad 20, \text{I.}$$

$$\Sigma\,(n_c l^2) = \Sigma\,(y^2)\,({\eta_{yx}}^2 - r^2) \quad 47, \textbf{XXVI.}$$

$$n\,\phi^2 = \sum \frac{(x - x')^2}{x'} \quad 45.$$

$$c = \sqrt{\frac{\phi^2}{1 + \phi^2}} \quad 45.$$

INDEX

Arabic numbers refer to pages.

area distribution, 26
accidental errors, 7

chance, 12
closest fit, line of, 36, 45
coefficient of correlation, 48
coefficient of mean square contingency, 63
contingency table, 53
contingency, total mean square, 63
correlation, chs. vi and vii
correlation, quality, 62
correlation coefficient, 48
correlation, partial, ch. vii
correlation ratio, 66
correlation, direct and inverse, 46

deviation, 9
deviation, standard, 20
direct correlation, ch. vi
distribution, area, 26
distribution, linear, 15
distribution, volume, 29
dot-diagram, 44

equations, area, 27
equations, normal, 34
equations, integral, 23, 27, 30
equations of type $y = ax^n$, $y = ab^x$, 37
equations, regression, 45
equations, volume, 29
error, 7
errors, accidental and systematic, 7
error, mean, 17
error of the interval, 72
error of last figure, 70
error of observation (correlation), 60
error, percentage-, graph, 22
error, personal, 7
error, probable, 20
error, probable of mean, 21
error, relative, 76
errors of functions, 76

fit, line of closest, 36, 44, 45

frequency, 8
functions, errors of, 76

graph, 'pencil-shot', 8
graph, percentage-error, 22
graph, 'school-boy', 9

interval, error of, 72
interval, 10
inverse correlation, 46

last figure, error of, 70
limits of correlation ratio and correlation coefficient, 67, 69; 49, 51
linear distribution, ch. iii
line of closest fit, 36, 44, 45
line, regression, 44

mean, 9
mean error, 17
mean square contingency, total, 63
mean square contingency, coefficient of, 63
mean, probable error of, 21
most probable point, 35

normal, equations, 34

observation, error of, 60
odds, 14

partial correlation, ch. vii
'pencil-shot' graph, 8
percentage-error graph, 22
personal errors, 7
probability, 12
probable error, 20
probable error of mean, 21
probable error of a correlation coefficient, 51
probable error of a correlation ratio, 69.
probable, most — point, 35

ratio, correlation, 66
relative error, 76

regression equations, 45
regression lines, 44

'school-boy' graph, 9
significance of a correlation coefficient, 50
standard deviation, 20

systematic errors, 7

table of contingency, 53
table, 2Y-, 23, 85, 86
total mean square contingency, 63

volume distribution, 29

Agricultural College and Research Institute Library
No. 2248
11 OCT 1950
Coimbatore

PRINTED IN GREAT BRITAIN AT THE UNIVERSITY PRESS, OXFORD
BY JOHN JOHNSON, PRINTER TO THE UNIVERSITY